

SAMPLING METHODS.
FOR
CENSUSES AND SURVEYS

Other books on Statistics
published by Charles Griffin & Co. Ltd.

By G. UDNY YULE *and* M. G. KENDALL
An Introduction to the Theory of Statistics

By M. G. KENDALL
The Advanced Theory of Statistics
(two volumes)
Rank Correlation Methods

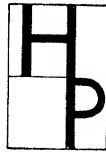
SAMPLING METHODS FOR CENSUSES AND SURVEYS

By

FRANK YATES, Sc.D., F.R.S.

HEAD OF THE DEPARTMENT OF STATISTICS, ROTHAMSTED EXPERIMENTAL STATION
AND OF THE RESEARCH STATISTICAL SERVICE OF THE AGRICULTURAL RESEARCH
COUNCIL AND THE MINISTRY OF AGRICULTURE • MEMBER OF THE UNITED NATIONS
SUB-COMMISSION ON STATISTICAL SAMPLING

PROPERTY OF
CARNEGIE INSTITUTE OF TECHNOLOGY
LIBRARY



LONDON
CHARLES GRIFFIN & COMPANY LIMITED
NEW YORK
HAFNER PUBLISHING COMPANY

1949

311.2

Y31

cop. 4

First published, 1949

All rights reserved

THE UNIVERSITY OF CHICAGO
LIBRARY

COPYRIGHT: CHARLES GRIFFIN & CO., LTD., LONDON

MADE IN GREAT BRITAIN

PRINTED BY BELL & BAIN LTD., GLASGOW

PREFACE

THIS book has been written primarily for those who have little or no previous training in mathematical statistics, but who have some training or experience in the presentation and handling of statistical data. It is consequently not written in the form of a mathematical treatise, and mathematical proofs have not been included. On the other hand an attempt has been made to cover all the modern developments of sampling theory which are of importance in census and survey work, and to give an adequate discussion of the complexities that are encountered in their practical application. This has necessitated fuller treatment of the subject than is to be found in textbooks on mathematical statistics, or than is normally included in statistical courses. Indeed, the orderly development imposed by the preparation of a book revealed a number of gaps in current theory which had to be filled in. Consequently the book should also prove of value to mathematical statisticians who are interested in sampling theory and its applications.

The work had its origin in a request of the United Nations Sub-Commission on Statistical Sampling, at their first session held at Lake Success in September, 1947, that a manual be prepared to assist in the execution of the projected 1950 World Census of Agriculture, and the 1950 World Census of Population. The Sub-Commission were particularly impressed with the need for a wider use of sampling in the less developed areas, and it was originally intended that only the sampling problems encountered in censuses and surveys in these areas should be dealt with. On reviewing the matter, however, I came to the conclusion that conditions differed so greatly in different areas that it would be necessary to cover a wide variety of methods, which in essentials differed little from the methods appropriate to censuses and surveys in more fully developed areas. It therefore seemed best to take the opportunity of writing a more general book. I believe that on balance the course taken will be advantageous to those concerned with censuses and surveys in the less developed areas, since the modern developments in sampling have been chiefly made in conjunction with its application in the more fully developed areas, and they can best be explained against the background of the material and problems to which they have been applied.

The various computational procedures have been illustrated, as far as practicable, by numerical examples. These examples in the main have an agricultural background, since this type of data was most readily accessible and is also particularly relevant to the original purpose of the book. For the most part the data on which they are based form a small part of the results

of much larger surveys. The examples do not in themselves serve as models for the reduction of large bodies of data, but once the general principles have been grasped no great difficulty should be found in planning this reduction, which presents very similar problems to those encountered in the analysis of material from complete censuses and surveys.

I have not attempted to ascribe priority in the discovery of particular methods. Indeed, such a task presents almost insuperable difficulties, since the methods used in many surveys are not at all fully reported, and the main developments have arisen chiefly through ingenious practical workers devising new methods of selection which seemed on commonsense grounds to be capable of giving specially accurate results, or appeared to possess other valuable properties.

The book has unfortunately taken longer to prepare than I had originally anticipated. This is mainly due to its widened scope and the various theoretical investigations that proved to be necessary. It has had to be prepared in considerable haste, and I can scarcely hope that it is entirely free from errors or omissions, or that perfect balance has been achieved. Certain minor discrepancies in the numerical work arise from the fact that some of the calculations have been carried out to more places of decimals than are finally reported, and that some of the standard errors have been computed on a slide rule. I shall be most grateful, however, if my attention is drawn to any errors of consequence that may be discovered.

I have been much helped by my wife in the planning of the book, and I have had considerable assistance in its preparation from various members of the Rothamsted Statistical Department. My thanks are particularly due to Dr. Rose O. Cashen and Mr. H. D. Patterson for computing and checking many of the examples, to Dr. P. M. Grundy and Mr. G. M. Jolly for their critical reading of the galley proofs, and to Miss Ruth Hunt and other members of the secretarial and computing staff for all their work and assistance. I also wish to thank the publishers and printers for the care taken in the preparation of this book in spite of the great haste that was necessary.

I have also received very considerable assistance from the discussions on the many and varied aspects of sampling that took place at the first two sessions of the United Nations Sub-Commission on Statistical Sampling, and from the Secretariat of the Statistical Office of the United Nations, who made available a very complete list of references on sampling on which the list given in this book is based.

F. YATES

ROTHAMSTED EXPERIMENTAL STATION,
7th February, 1949.

CONTENTS

CHAPTER 1

THE PLACE OF SAMPLING IN CENSUS WORK

	PAGES
1.1 The sampling process	1
1.2 Sampling errors	2
1.3 The place of sampling in census and survey work	2
1.4 Development of the use of sampling in censuses and surveys	5
1.5 Method of presentation	6
1.6 Terminology and notation	7

CHAPTER 2

REQUIREMENTS OF A GOOD SAMPLE

2.1 Bias	9
2.2 Methods of selection which give rise to bias	9
2.3 Avoidance of bias in selection	10
2.4 Examples of biased selection	11
2.5 Bias arising from faulty demarcation of the sampling units	15
2.6 Bias in estimation	16
2.7 Circumstances in which bias is permissible	17
2.8 Methods of reducing the random sampling error	17
2.9 Choice of unit	19

CHAPTER 3

THE STRUCTURE OF VARIOUS TYPES OF SAMPLE

3.1 Definition of frame and sampling unit	20
3.2 Random sample	21
3.3 Stratification with uniform sampling fraction	23
3.4 Multiple stratification	25
3.5 Stratification with a variable sampling fraction	28
3.6 Systematic samples from lists	29
3.7 An example of alternative ways of sampling highly variable material	30
3.8 Multi-stage sampling	34

CONTENTS

	PAGES
3.9 Sampling with probabilities proportional to size of unit ...	35
3.10 Sampling from within strata with probabilities proportional to size of unit	36
3.11 An example of sampling by administrative areas	36
3.12 Multi-phase sampling	38
3.13 Balanced samples	39
3.14 Systematic samples from areas	41
3.15 Line sampling	42
3.16 The principle of the moving observer	43
3.17 Interpenetrating samples	44
3.18 Sampling on successive occasions	45
3.19 Composite sampling schemes	46
3.20 Combination of complete census and sample survey	47

CHAPTER 4

PRACTICAL PROBLEMS ARISING IN THE PLANNING OF A SURVEY

4.1 Questions requiring consideration	48
4.2 Definition of the population	49
4.3 Determination of the details of the information to be collected	51
4.4 Inter-relations of groups of natural units	54
4.5 Practicability of obtaining the required information	55
4.6 Methods of collecting the information	57
4.7 Methods of dealing with non-response	59
4.8 The frame	60
4.9 Frames suitable for censuses and surveys of human population	62
4.10 Frames from lists of individuals	63
4.11 Frames from complete population censuses	65
4.12 Frames from lists of households or dwellings	66
4.13 Frames provided by town plans	68
4.14 Frames provided by maps of rural areas	69
4.15 Frames from lists of villages	70
4.16 The 1946 population sample for Greece	71
4.17 Master samples	73
4.18 Localized population surveys	75
4.19 The U.S. series of employment estimates	76

CONTENTS

	PAGES
4.20 Frames suitable for special classes of a human population ...	78
4.21 Frames suitable for the survey of economic institutions ...	79
4.22 Market research and opinion surveys	79
4.23 Frames for agricultural censuses and surveys	81
4.24 Use of maps as frames in agricultural surveys	82
4.25 The 1942 Census of Woodlands	83
4.26 Frames for undeveloped areas	85
4.27 Use of aerial survey photographs	86
4.28 Crop estimation	87
4.29 Estimation of yield by the harvesting of sample areas	90
4.30 Crop forecasting	92
4.31 Determination of the size of sample when the sample is fully random	94
4.32 Some general rules on size of sample	97
4.33 Pilot and exploratory surveys	99

CHAPTER 5

PROBLEMS ARISING IN THE EXECUTION AND ANALYSIS OF A SURVEY

5.1 Types of problem	102
5.2 Administrative organization	102
5.3 Design of forms	103
5.4 Special tests of questionnaires and investigators	105
5.5 Selection, training and supervision of field investigators	105
5.6 Control of the accuracy of the field work	106
5.7 Arrangements for follow-up in the case of non-response	107
5.8 Statistical analysis	108
5.9 Methods of handling the data	109
5.10 Cope-Chat cards	110
5.11 Punched cards	112
5.12 The sorter and sorter-counter	113
5.13 The tabulator	113
5.14 The reproducing summary punch	116
5.15 The multiplying punch	117
5.16 The collator	118
5.17 Systems of coding for punched cards	118

CONTENTS

	PAGES
5.18 Arrangement of information on punched cards	120
5.19 General remarks on the planning of the computations	123
5.20 Control of numerical accuracy in the analysis	124
5.21 The use of sampling in the statistical analysis	128
5.22 Adjustment of the results to compensate for defects in the sample	129
5.23 Critical analysis of survey data	131
5.24 Method of fitting constants	137
5.25 Preparation of reports	141

CHAPTER 6

ESTIMATION OF THE POPULATION VALUES

6.1 Possibility of alternative estimates	145
6.2 Notation	146
6.3 General rules	147
6.4 Random sample	148
6.5 Stratified sample with uniform sampling fraction	150
6.6 Random sample, stratified after selection	152
6.7 Stratified sample with variable sampling fraction	153
6.8 Use of supplementary information in estimation	155
6.9 Ratio method: random sample	159
6.10 Ratio method: stratified sample with uniform sampling fraction	160
6.11 Ratio method: stratified sample with variable sampling fraction	161
6.12 Regression method: random sample	162
6.13 Regression method: stratified sample with uniform sampling fraction	164
6.14 Regression method: stratified sample with variable sampling fraction	164
6.15 Use of regression to calibrate eye estimates	165
6.16 Sampling with probabilities proportional to size of unit	167
6.17 Sampling from within strata with probabilities proportional to size of unit	169
6.18 Multi-stage sampling, no supplementary information	170
6.19 Multi-stage sampling with supplementary information	171
6.20 Systematic and balanced samples	174
6.21 Sampling on two successive occasions	175
6.22 Sampling on a number of successive occasions	179

CONTENTS

CHAPTER 7

ESTIMATION OF THE SAMPLING ERROR

	PAGES
7.1 Sampling errors of a random sample	183
7.2 Sampling from a finite population	187
7.3 The normal law of error	190
7.4 Qualitative variates	193
7.5 Standard errors of functions of estimates	196
7.6 Stratified random sample with possibly unequal variances within strata	201
7.7 Pooled estimate of error: the analysis of variance	205
7.8 Ratio method: random sample	212
7.9 Ratio method: stratified sample with uniform sampling fraction	215
7.10 Ratio method: stratified sample with variable sampling fraction	216
7.11 Ratio method: integral values of the supplementary variate ...	217
7.12 Regression method: random sample	218
7.13 Regression method: stratified and balanced samples	221
7.14 Calibration of eye estimates	222
7.15 Sampling with probabilities proportional to size of unit ...	224
7.16 Sampling from within strata with probabilities proportional to size of unit	225
7.17 Multi-stage sampling	226
7.18 Systematic samples	229
7.19 Sampling on successive occasions	233
7.20 The error graph	235
7.21 Sub-sampling for the estimation of error	238
7.22 Rounding-off and grouping errors	238
7.23 Determination of errors due to bias	239
7.24 Interpenetrating samples: comparison of observers	241
7.25 Estimation of the sampling error from duplicate samples ...	242
7.26 Presentation of sampling errors in extensive surveys	243

CHAPTER 8

EFFICIENCY

8.1 General remarks	246
8.2 Qualitative data	248
8.3 Random sample and stratified sample with uniform sampling fraction	249

CONTENTS

	PAGES
8.4 Multiple stratification	254
8.5 Stratified sample with variable sampling fraction	254
8.6 Supplementary information	256
8.7 Two-phase sampling	258
8.8 Sampling on successive occasions	260
8.9 Sampling with probability proportional to size of unit	262
8.10 Interpretation of the analysis of variance	266
8.11 Multi-stage sampling	268
8.12 An example of a pilot sampling scheme for crop estimation	273
8.13 A special case of two-stage sampling	278
8.14 Effect of change in size of the sampling units	279
8.15 Variation in size of strata	280
8.16 Efficiency in terms of cost	283
8.17 Minimization of the cost function	285
8.18 Losses due to errors	292
8.19 Concluding remarks	294

APPENDIX TABLES

A.1 Random numbers	297
A.2 The normal distribution	298
 BIBLIOGRAPHY	 299
 INDEX	 313

LIST OF EXAMPLES

Methods of random selection	...	...	...	Ex. 3.2.a, 3.2.b, 3.2.c
Stratification without control of sub-strata :				
selection of sample	...	...	...	Ex. 3.4
Hertfordshire farms :				
random sample	...	...	...	Sect. 3.7, 4.31, Ex. 6.6, 6.9.a, 6.10, 6.12.a, 7.2.b, 7.6.b, 7.8.a, 7.9, 7.12.a, 8.3.b, 8.6.c
stratified random sample	...	...	...	Sect. 3.7, Ex. 6.5, 7.6.a, 8.2.b, 8.3.a
variable sampling fraction	...	...	...	Sect. 3.7, Ex. 6.7, 7.7.a, 8.3.c, 8.5.a, 8.5.b
samples of parishes	...	...	...	Sect. 3.11, Ex. 6.17, 7.16, 8.9.b, 8.9.c
two-stage samples of parishes and farms	...	...	...	Sect. 3.11, Ex. 8.11.a, 8.11.b, 8.17.c
Family income in Norfolk-Portsmouth, Virginia				Sect. 4.31, Ex. 7.1.b, 7.22
Potatoes : yields of varieties in different regions				Sect. 5.23, 5.24, Ex. 7.5, 7.7.b
National Farm Survey	...	...	...	Sect. 5.21, Ex. 6.8
Housing survey : percentage defective	...	...	...	Ex. 6.4.a, 7.4
Sample from a normal distribution	...	...	...	Ex. 6.4.b, 7.1.a, 7.2.a, 7.3
Population sample in Southern Rhodesia	...	...	...	Ex. 6.9.b, 7.8.b
Census of Woodlands : volume of timber	...	...	...	Ex. 6.12.b, 7.12.b, 8.6.b, 8.17.a
Eye estimation of wheat	...	...	...	Ex. 6.15, 7.14
Point sampling for area and yield of a crop	...	...	...	Ex. 6.16.a, 6.16.b, 7.15, 8.17.b
Survey of Fertilizer Practice	...	...	...	Ex. 6.19, 7.17, 7.23, 8.6.a, 8.9.a, 8.17.d
Sampling on two successive occasions : percentage of solids-not-fat	...	...	...	Ex. 6.21, 7.19.a, 8.8

LIST OF EXAMPLES

Sampling on six successive occasions : percentage of solids-not-fat	Ex. 6.22, 7.19.b, 8.8
Census of Woodlands : area of woodland from maps	Ex. 7.18
Census of Woodlands : interpenetrating samples	Ex. 7.25
Effect of stratification on estimation of an attribute	Ex. 8.2.a
Point sampling and area measurements of fields	Ex. 8.11.c
Crop estimation scheme for wheat yield ...	Sect. 8.12

SAMPLING METHODS FOR CENSUSES AND SURVEYS

CHAPTER 1

THE PLACE OF SAMPLING IN CENSUS WORK

1.1 The sampling process

Sampling, that is, the selection of part of an aggregate of material to represent the whole aggregate, is a long-established practice. Simple examples are provided by a handful of grain taken from a sack, or a piece of cloth cut off a roll. In these cases little attention need be paid to the selection process, since the whole of the material is similar or well-mixed, and any part of it if not too small is likely to be closely representative of the whole. When, however, the aggregate to be sampled consists of units which are somewhat dissimilar amongst themselves, and which are not well-mixed, a small sample of these units may not be representative of the whole aggregate. Even if units are selected from different parts of the aggregate, and other suitable precautions are taken, the sample is likely to a certain extent to be unrepresentative owing to the chance inclusion of an undue proportion of units of a particular type. It will clearly not be representative if units of a particular type are chosen deliberately to the exclusion of other types, or if the process of selection is such that certain types of unit are favoured at the expense of others. Thus in sampling a heap of coal by taking a few shovelfuls from the edges, too great a proportion of the large lumps will be obtained, since the large lumps tend to roll down the sides and be distributed round the edges of the heap. Similarly in the sampling of continuous material, a single portion, even if quite large, may not be adequately representative; a piece of cloth cut off the end of a roll in which the quality of the weaving varies progressively, will not form an adequate sample of the whole roll.

Census and survey work is normally carried out on material made up of dissimilar units. Censuses of population, censuses of industrial production, and censuses of agriculture have the common feature that the aggregate of material embraces a large number of separate units which are often markedly dissimilar in various respects. In many cases the purposes for which the information is required are adequately served if a proportion only of the units are covered, but because of the dissimilarity of the different units neither

haphazard nor casual selection, and still less deliberate selection, can be expected to provide a representative sample. Rigorous processes of selection have therefore to be used.

Censuses carried out on a properly selected sample will be called *sample censuses*. There has in the past been a tendency to use the term sample to refer to the results of an attempted complete census in which there has been failure to obtain information from a substantial proportion of the units. Its use in this sense is strongly to be deprecated; instead the term *incomplete census* is suggested. The term *sample* should be reserved for a set of units or portion of an aggregate of material which has been selected in the belief that it will be representative of the whole aggregate.

1.2 Sampling errors

Whether or not a sample will give results which are sufficiently representative of the whole aggregate depends primarily on whether the errors introduced by the sampling process are sufficiently small not to invalidate the results for the purposes for which they are required. Even if a proper process of selection is employed, the sample cannot be exactly representative of the whole aggregate. The inevitable errors which then occur in the results are termed the *random sampling errors* of these results. The average magnitude of these random sampling errors will depend on the size of the sample, on the variability of the material, on the sampling procedure adopted, and on the way in which the results are calculated.

It is a fortunate fact that if a proper process of selection is adopted, the average magnitude of the random sampling errors, and indeed the expected frequency of occurrence of errors of any magnitude, can be calculated from the detailed results obtained from an actual sample. The methods by which this can be done depend on the mathematical theory of statistical sampling.

An extension of the analysis involved in the calculation of these errors enables the relative accuracy of the different sampling methods which can be employed on the same material to be assessed, and thus enables further surveys to be more efficiently planned.

It is the development of these processes that has changed sampling from a speculative and uncertain procedure to a method having definite and determinable precision. Sampling has thus become a reliable method in which full confidence can be placed. In addition, the possibility of setting ascertainable limits to the random sampling errors has served to throw into prominence those other types of error which arise from faulty selection processes or faulty methods of observation, or which exist in some other source of information with which the sampling results are being compared.

1.3 The place of sampling in census and survey work

Sampling will only be of use in census work if, as mentioned in Section 1.2, the sampling errors are sufficiently small not to affect the validity of the results

for the purposes for which they are required. This will in part be a function of the degree to which the results have to be broken down. If only overall results for the whole population are required, a given degree of accuracy will be attained with a far smaller sample than will be the case if detailed results for different parts of the population (*e.g.* different regions, towns, etc.) are required. In certain circumstances the sample may have to be so large that there will be little point in using a sample census in place of a complete census. Obviously, in the extreme case where information on all the individual units is required, this can only be obtained by a complete census.

Another factor which influences the decision whether or not to use sampling is the relative difficulty and cost of organizing a sample census and a complete census. The amount of effort and expense required to collect information is always greater per unit for a sample than for a complete census. In addition a sample census presents its own organization problems, some of which are absent from a complete census, and it occasionally happens, if the information required is very simple, that a complete census can be carried out through the ordinary administrative channels, whereas a sample census requires the setting-up of a separate organization. Usually, however, if the size of the sample needed to give the required accuracy represents only a small fraction of the whole population, the total effort and expense required to collect the information by sampling methods will be very much less than that required for a census of the whole population.

In many cases, therefore, sampling results in great economy of effort. It has also other advantages which are not so immediately apparent. In the first place, the completeness and accuracy of the returns may be much more easily ensured if the information is collected from only a small proportion of the population. If, for example, questionnaires are sent through the post, it is frequently impossible in a complete census to bring pressure to bear on those who fail to make their returns, even where the completion of these questionnaires is compulsory, owing to the large numbers of individuals involved. In the case of a sample, the smaller number of individuals enables follow-up notices to be sent and telephone calls and visits to be made. The separate returns can also be much more carefully scrutinized, and further enquiries undertaken where there is reason to doubt their accuracy.

Secondly, it is possible to obtain more detailed information in a sample census. Although the burden on the individual of furnishing more detailed information is not lessened, except when different items of information can be obtained from different individuals, the individuals concerned are more likely to be willing to provide such information if they know that they represent a small sample of the whole population. Detailed information, when obtained, can be more easily handled, both at the stage of abstraction and coding of the original information and in the analysis of the coded results. Owing to the reduced volume of material that has to be handled the quality of the abstraction and analysis can also be improved, the former because a higher grade of clerical labour can be employed, with better supervision, and the latter because the

data can be classified in many more ways with the same amount of computing or machine time.

Thirdly, in many types of census the use of sampling makes possible a very considerable increase in speed, both in the execution of the field work, and in the analysis of the results. Speed in analysis can also be obtained, in the case of a complete census, by taking a sample of the returns for abstraction and analysis. This device is frequently of value for providing preliminary results quickly, even when a final analysis of the whole of the returns is ultimately required.

The use of sampling is essential for investigations of the sociological type in which extensive and detailed information has to be collected from individuals, many of whom have neither the education nor experience required to answer detailed questionnaires without assistance. It is equally essential in investigations requiring skilled physical observations and measurements. Such investigations can only be carried out by the use of trained investigators, and complete investigations covering any large group of the population or body of material are consequently impossible, both on grounds of expense and because, even if the expense can be tolerated, a sufficient body of investigators can rarely be recruited and trained.

For an investigation of this kind involving the collection of elaborate information the term *survey* is usually employed. It seems a mistake, however, to confine the word *survey* to a sample survey or the word *census* to a complete census. Thus B. Seebom Rowntree (1901), when he carried out an investigation into the social and economic conditions of all working-class families in York, correctly described this as a survey.

Although the use of sampling necessarily introduces certain inaccuracies, owing to sampling errors, the results obtained by sampling are frequently more accurate than those obtained in a complete census or survey. The random sampling errors are always assessable. The other errors to which a survey is subject, such as incompleteness of returns and inaccuracy of information, are liable to be very much more serious in a complete census than in a sample census, since far more effective precautions can be taken, to see that the information is accurate and complete in a sample census. Furthermore, the use of sampling greatly facilitates the imposition of additional more detailed checks. Indeed, a complete census can only be properly tested for accuracy by some form of sampling check.

On the other hand, the claim that is sometimes made that the reliability and accuracy of the results of a properly planned sample census can be assessed with full objectivity from the results themselves is only partly true. The random sampling errors can be so assessed, and under certain circumstances it is possible to obtain comparisons between different investigators. If all investigators or respondents tend to make the same kind of error, however, this will not be revealed in the results, whether the census is complete or carried out on a sample.

In respect of coverage a sample census may in certain circumstances be

less reliable than a complete census. It is, for example, relatively simple for an investigator to ascertain by direct question whether an individual has already been included in a population census, and simple intensive checks of certain areas, say villages, can be made in a similar manner to verify that there is no appreciable number of omissions. Similarly, in a survey of physical objects such as houses, a marking system or other suitable device can often be used to guard against duplication and omission. Such checks are impossible in the case of a sample census.

This is one of the most difficult points in the practical design of many sample surveys, particularly in undeveloped areas. To overcome it complete enumeration can sometimes be used in conjunction with sampling. Where a complete enumeration of the whole of the population or aggregate of material presents no particular difficulty, but where the collection of detailed information from all units would be a difficult or impossible undertaking, a complete enumeration can be carried out. This is then used as a basis for the selection of the sample for such sample censuses and surveys as are required to provide the detailed information.

1.4 Development of the use of sampling in censuses and surveys

Prior to the development of the appropriate methods of estimation of sampling errors and a clear recognition of the conditions governing satisfactory methods of selection of the sample, the use of sampling in census and survey work often proved unsatisfactory. There are many early examples of sample censuses and surveys which are defective in one way or another. Even when the basic principles of the simpler forms of sampling were understood, the attempted use of more complicated forms before methods of evaluating their errors and relative efficiency had been worked out gave rise to further defective surveys.

This has led to a certain mistrust of sampling, which still exists in some quarters. During recent years, however, there has been a rapid growth in the use of sampling in various countries. This development has been greatly stimulated by the war and its attendant measures of large-scale economic control. Such measures, if they were to be effective in the changing conditions met with in wartime, demanded an efficient and speedy information service which only the sampling method could supply. This has resulted in further improvements in technique through the stimulation of research into the theory of sampling methods, and the provision of basic data for practical investigations of the relative efficiency of the various methods in different fields.

It still remains true, however, that in inexperienced hands sampling may give unsatisfactory results, owing to the use of faulty methods of selection, inappropriate sampling design, or inefficient methods of estimation. The prime requirement of any large-scale sample survey is therefore that the organization of the survey should be carried out by a person who has adequate knowledge and experience of sampling methods and their application. The methods

employed must be thoroughly sound, theoretically and practically, both in order that satisfactory results may be ensured, and also in order that mistrust cannot subsequently be engendered by criticism of the methods adopted. It must never be forgotten that it is not sufficient to provide results which are in fact correct. They must also be generally accepted if they are to have their full value.

It is sometimes stated that no large-scale sample census or survey should be carried out without the advice of an expert mathematical statistician with experience of such work. Unquestionably, if the services of such an expert can be secured this is all to the good, but my own experience is that no one expert can be expected to supervise adequately more than a very few surveys at any one time, since adequate supervision demands a very full knowledge both of the material that is to be surveyed and of local conditions, coupled with close attention to detail at all stages. An expert acting in an advisory capacity is therefore no substitute for the statistician on the spot, who must be prepared to accept responsibility for the planning, execution and analysis of the survey. To do this he must himself have both an adequate knowledge of sampling procedure and thorough knowledge of the material and local conditions.

Consequently, if full and effective use is to be made of sampling methods, statisticians and others who already have experience of the conduct of complete censuses but no training in sampling methods must themselves undertake a study of these methods, in order that they may decide in what ways these can be applied to their own problems. The function of the expert then becomes one of advice on exceptional problems, rather than one of detailed supervision.

Fortunately the principles underlying good sampling methods are not unduly difficult to understand, and provided a proper respect is observed for the fundamental rules of procedure I believe they can be successfully applied by those who have statistical experience but who are not primarily mathematical statisticians.

1.5 Method of presentation

The method of presentation adopted in this book is to take the various parts of the sampling process in roughly the order they are encountered in the execution of a census or survey, and discuss the various aspects of each part in turn. Thus Chapters 2 and 3 describe the various types of sample that can be used, and the general principles to be followed in the selection of a sample, Chapter 4 deals with the practical planning of a survey, and Chapter 5 with the problems encountered in its execution and in the abstraction of the results. The remaining chapters are concerned with the more strictly statistical problems. Chapter 6 deals with the various methods of estimating the population values, Chapter 7 with the estimation of sampling errors, and Chapter 8 with the determination of the relative efficiency of the various sampling methods. This method of presentation has the advantage that the

more practical aspects of sampling procedure are dealt with first. It is true that knowledge of the statistical techniques described in the later chapters is necessary before the relative merits of different methods of sampling any particular type of material can be accurately assessed. The detailed application of these techniques, however, is the province of those responsible for the numerical analysis of the results, whereas the planning is also the concern of those who require the information and those who are concerned with its collection. The planning can be undertaken much more efficiently, and with added interest, if all concerned understand in general terms the underlying problems. It is hoped that study of the first five chapters will give this understanding. If they also act as a stimulus to the study of Chapter 6 and the first few sections of Chapter 7 the understanding should be correspondingly deepened.

For those responsible for the numerical analysis of the results, and for the assessment of the relative efficiency of the different possible methods, thorough study of the whole book is necessary. This study should include the reworking of the numerical examples. Only by this procedure can a thorough grasp of the details of the various methods be obtained.

The separation of the discussion of the methods of estimation of the population values, of the sampling errors, and of efficiency necessarily involves a good deal of cross-reference, particularly in the numerical examples. Since this appeared inevitable, it was with some hesitation that the chosen method of presentation was adopted. On balance, however, this disadvantage appeared to be outweighed by the advantage of being able to present as a whole the relatively simple techniques involved in estimation before the more complicated techniques required for the estimation of error and the assessment of relative efficiency. It is believed that this will make the book more useful to those who do not require to go deeply into these latter techniques. For those who prefer it, there is nothing to prevent the simultaneous study of the corresponding sections of Chapters 6 and 7, or indeed of Chapters 6, 7 and 8. Chapters 6, 7 and 8 may also, if desired, be taken before Chapters 4 and 5.

1.6 Terminology and notation

The question of terminology was considered by the United Nations Sub-Commission on Statistical Sampling, at its second session held in Geneva in September, 1948. Their recommendations are included in a memorandum entitled *Recommendations concerning the Preparation of Reports on Sampling Surveys*. With a few minor exceptions the terminology adopted in this book is that recommended by the Sub-Commission.

New conventions have been adopted for the mathematical notation. The use of bold face and Gill Sans type for population values and their estimates, and of capital letters for the population totals, has enabled the formulæ to be presented in a very simple, and it is hoped easily understandable form. By the use of this notation the elaborate summation notation which has become

current in much of the literature on sampling has been avoided. It is recognized that the notation is not particularly convenient for manuscript and typescript, but the difficulty can in fact be overcome by the use of single and double underlining, with the corresponding verbal descriptions of "sub-bar" and "sub-double-bar."

CHAPTER 2

REQUIREMENTS OF A GOOD SAMPLE

2.1 Bias

The principal object of any sampling procedure is to secure a sample which, subject to limitations of size, will reproduce the characteristics of the population, especially those of immediate interest, as closely as possible.

At first sight it might appear that the most accurate results could be obtained by deliberate selection of the units to be included in the sample. In particular, if averages only are of interest, units might be selected which appear to be nearest to the average. If, for example, a quick assessment of the yield per acre of an agricultural crop is required, district officers might be asked to select some "average" fields in each district, and to determine the yields of these fields.

Such a sample is unfortunately very often of little value. Its primary fault is that it may well be *biased*, that is, the selection of all the fields may be affected by similar errors. Thus, in order to enhance the reputation of their districts, all district officers may tend to select fields which yield more heavily than the average, or, if they feel that the interests of the farmers or the country may be furthered by an underestimate, they may select fields which yield less than the average.

Even if the district officers can be trusted to be completely objective, considerable unconscious errors of judgment, all tending in the same direction, may still occur, and such errors may far outweigh any increase in accuracy resulting from deliberate selection. Nor will increase in the number of officers concerned in the selection necessarily improve matters, since all may be subject to the same type of error.

We may consequently distinguish between two types of sampling error, those arising from biases in selection, etc., and those due to chance differences between the members of the population included in the sample and those not included. The aggregate of the former in the sample will be termed the *error due to bias* and the aggregate of the latter the *random sampling error*, or when bias is known to be absent, the *sampling error*. The total sampling error will, of course, be made up of the bias, if any exists, and the random sampling error. The essence of bias is that it forms a constant component of error which does not decrease, in a large population, as the number in the sample increases, whereas the random sampling error decreases on the average as the number in the sample increases.

2.2 Methods of selection which give rise to bias

There are a number of ways in which faulty selection of the sample may give rise to bias. The main causes may be broadly classified as follows:—

- (1) Deliberate selection of a "representative" sample. This is the type of bias described above.

- (2) A procedure of selection depending on some characteristic which is correlated with properties of the unit which are of interest. Many haphazard selection processes give rise to biases of this kind.
- (3) Conscious or unconscious bias in the selection of a "random" sample. If a proper random process is not strictly adhered to, the investigator, although claiming that his sample is random, may allow his desire to obtain a certain result to influence his selection. This type of bias is particularly serious, since its existence may not be immediately apparent.
- (4) Substitution. Investigators often substitute another convenient member of the population when difficulties are encountered in obtaining information. Thus, in a house-to-house survey the next house may be taken when there is no reply. This will necessarily lead to a preponderance of houses of the type that are occupied all day, *e.g.* houses of people with families.
- (5) Failure to cover the whole of the chosen sample. If no second visit is made to houses from which no reply is received there will still be bias even though no substitution is attempted. This fault is particularly prevalent in postal questionnaires, which are often very incompletely returned. Returns are clearly likely to be received from individuals who are specially interested in the objects of the survey, or possess other characteristics which make them unrepresentative of the whole population.

2.3 Avoidance of bias in selection

It is clear that, if possibilities of bias exist, no fully objective conclusions can be drawn from a sample. The first essential of any sampling procedure must therefore be the elimination of all important sources of bias.

The simplest, and the only universally certain way, of avoiding bias in the selection process is for the sample to be drawn either entirely at random, or at random subject to restrictions which, while improving the accuracy, are of such a nature that they do not introduce bias into the results. In some cases, however, certain forms of *systematic* selection, such as the selection of names at equal intervals down a list, or the use of an evenly spaced grid of points on a map, may be permissible.

Random selection does *not* mean haphazard selection. A random sample can only be obtained by adherence to some proper random process, such as the drawing of lots or the use of a table of random numbers. Sticking pins into a map will not give a random distribution of points in a map. The selection of houses by walking through the streets of a town will not give a random selection of houses in the town. The words "random" and "random sample" are, in fact, gravely abused. For this reason, if for no other, the method of selecting the sample should be specified in all accounts

of the results of sample surveys and censuses, and indeed in all sampling work.

In order to prevent careless or deliberately biased selection on the part of investigators it is often important in large-scale work for the selection to be done in some central office, in such a manner that no element of choice is left to the investigators, and in such a manner also that checks on the field work can be imposed if necessary. Even in cases in which a less rigorous method of selection may be judged to be satisfactory, it may be necessary to impose a rigorous method in order to prevent criticism on this ground by those not familiar with the details of the work.

2.4 Examples of biased selection

It may be well at this stage to give some actual examples of cases in which an unsatisfactory method of selection has introduced serious bias into the results.

The first example is taken from a paper by Kiser (1934, D). A sample of households was taken in Syracuse, U.S.A., in 1930 and 1931, with the object of making a study of morbidity. It was also intended to use this sample for the study of birth-rates. Before beginning this latter study, which was subsidiary to the morbidity study, a comparison was made of the sizes of households of the sample with those of the corresponding census tracts. This comparison is shown in Table 2.4.a. (Households of one were not included in the survey.)

TABLE 2.4.a.—SAMPLE OF HOUSEHOLDS IN SYRACUSE: DISTRIBUTION OF HOUSEHOLDS ACCORDING TO SIZE, IN THE ORIGINAL SAMPLE, AND IN THE CENSUS TRACTS

Number in household	Original sample		Census tracts	
	Number	Per cent.	Number	Per cent.
2	254	19.4	1,762	26.8
3	338	25.9	1,745	26.5
4	307	23.5	1,438	21.9
5	201	15.4	853	13.0
6	106	8.1	388	5.9
7	46	3.5	208	3.2
8	25	1.9	96	1.5
9 and over .	29	2.2	86	1.3
TOTAL . . .	1,306	99.9	6,576	100.1

It is immediately apparent from the table that the sample contains a considerably greater proportion of large households than exist in the whole population. Households of two are under-represented in the sample to the

extent of 7.4 per cent. of all households, or 28 per cent. of the households of this size. This deficiency is attributed by Kiser to the failure of enumerators to revisit missed households, in which childless married women working away from home are likely to predominate. In order to provide a more satisfactory sample it was necessary to make a further survey of those families that were missed altogether at the time of the morbidity survey.

It is interesting to note that the sample was apparently considered satisfactory for the morbidity study, as is indicated by the statement that the workers "had been primarily concerned with securing a sample representative of the area in regard to prevalence of sickness rather than size of household." Actually such a biased sample can scarcely be regarded as wholly satisfactory for even a morbidity study, since sickness rates are likely to vary with the size and composition of the family.

The second example is one obtained at Rothamsted in an experimental sampling of a collection of stones (Yates, 1936, *b*, H). The stones, a number of flints of varying sizes, some 1200 in all, were spread out on a table, and twelve observers were each instructed to choose three samples of twenty stones which should represent as nearly as possible the size distribution of the whole collection. Table 2.4.b gives the mean weights per stone of these 36 samples, and also the true mean weight of the whole collection.

TABLE 2.4.b—MEAN WEIGHT PER STONE IN SAMPLES OF 20 STONES (oz)

Observer	1	2	3	4	5	6	7	8	9	10	11	12
Sample 1	1.9	2.4	2.4	1.9	2.2	2.8	2.4	1.6	2.2	2.6	2.4	2.4
Sample 2	1.8	3.0	2.4	2.0	2.7	2.6	2.6	2.0	2.2	2.2	2.4	3.0
Sample 3	1.7	2.4	2.1	2.0	3.1	2.8	2.5	2.0	2.2	3.1	1.8	2.4
Mean	1.8	2.6	2.3	2.0	2.7	2.7	2.5	1.9	2.2	2.6	2.2	2.6

Mean of all samples: 2.34 oz. True mean: 1.91 oz.

It is apparent that there is a tendency, which is common to most observers, to select stones which are on the average larger than those of the whole collection. Of the twelve observers ten chose samples whose mean weight was above the mean weight, 1.91 oz., of all the stones, the mean for all samples being 2.34 oz. This tendency is consistent from sample to sample. Thus, of the thirty samples chosen by the above ten observers, all but two had mean weights greater than the mean weight of all stones, while all three samples of observer 1 were less than the correct mean.

In this example the selection was deliberate. A further example showing similar effects arising from haphazard selection (claimed by the observer to

be "random") is provided by some observations obtained in the course of a scheme of sampling observations on the growth of wheat instituted by the Agricultural Meteorological Committee (Yates, 1935, A).

In this scheme measurements on the heights of shoots of wheat were made at regular intervals on observation plots at a number of centres. A detailed procedure had been laid down for the random location on each occasion of 128 quarter-metre lengths of row in sets of 4 on contiguous rows. The height measurements were made on the 256 shoots at the ends of these lengths—test observations conducted at another time indicated that this method of selection was virtually random. At one centre a drill with fewer rows than normal had to be used, and as a result only 192 shoots were available for measurement on each occasion. In order to provide the number of observations laid down and thereby, as he thought, improve his results, the observer selected "at random" two additional shoots from each set of three quarter-metre lengths. Fortunately he booked the observations on these additional shoots separately.

Figures 2.4.a and 2.4.b show the distribution of the regular and additional measurements taken on the 31st May and on the 28th June respectively. The deviations from the set means of the regular measurements are shown. Suitable adjustments, details of which are given in the original paper, are made to the additional measurements to give fair representation of the variability as well as the bias in the mean.

Examination of Figure 2.4.a indicates that on this date the additional measurements show a considerable preponderance of positive deviations with a corresponding deficiency of negative deviations. There is, in fact, a tendency to select shoots which are higher on the average than those of a truly random sample, the difference in the average height being $+ 3.3$ cm. This difference is clearly in the nature of a bias, and cannot be attributed to random sampling errors.

The situation was entirely different on the 28th June, as is shown in Figure 2.4.b. At this date the deviations of the additional measurements, both positive and negative, are smaller on the average than are those of the regular observations; in other words, there is a tendency to select shoots which are nearer the mean height than they would be on the average in a truly random sample. In spite of this, there is again a considerable bias, this time negative, the mean difference being $- 2.7$ cm. In this case, therefore, a single additional shoot will give a value which on the average is closer to the true mean value than is the value given by a single randomly located shoot, but as the number of shoots is increased the relative accuracy of the random sample progressively increases, and with the numbers of shoots actually taken, the random sample is considerably more accurate.

This example provides an illustration of a case where the biases on the two occasions, though arising from similar defects in selection, are of very different magnitude, and indeed of opposite sign. Consequently the difference of the two sets of measurements will also be seriously affected by bias. In

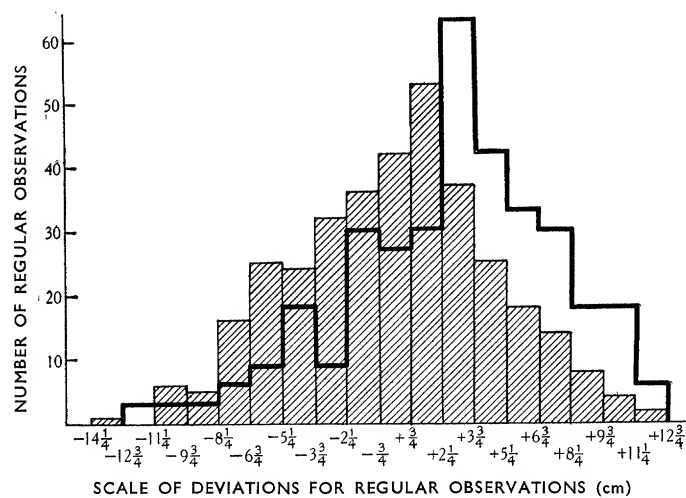


FIG. 2.4.a—DISTRIBUTION OF REGULAR OBSERVATIONS (*shaded*) AND ADDITIONAL OBSERVATIONS (*unshaded*) OF HEIGHTS OF WHEAT SHOOTS ON 31st MAY

(By courtesy of the editor of the *Annals of Eugenics*.)

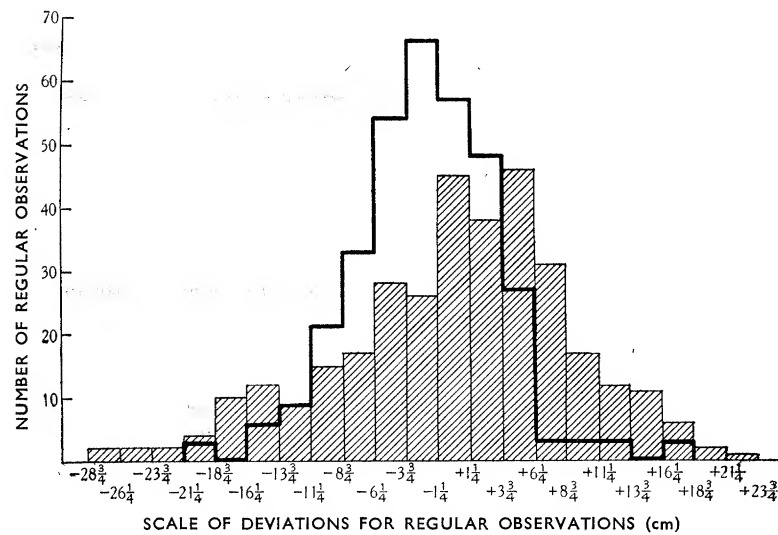


FIG. 2.4.b—DISTRIBUTION OF REGULAR OBSERVATIONS (*shaded*) AND ADDITIONAL OBSERVATIONS (*unshaded*) OF HEIGHTS OF WHEAT SHOOTS ON 28th JUNE

(By courtesy of the editor of the *Annals of Eugenics*.)

this case the growth rate of the wheat would have been underestimated by nearly 10 per cent. had only the additional measurements been available.

These biases are, of course, of the type that might be expected. When the shoots are only half-grown and there is nothing much to be seen except the top leaves there will be a tendency to pick the longer shoots, but when the crop has come into ear the observer can see shoots of all lengths, and is more likely to select shoots somewhere near the average, omitting both very long and very short shoots. The strong negative bias of the last set of measurements shows that this selection was not particularly effective in improving the accuracy of the sample.

2.5 Bias arising from faulty demarcation of the sampling units

Any consistent errors in measurement will clearly give rise to bias, whether the measurements are carried out on a sample or on all the units of the population. The danger of such errors is, however, likely to be greater in sampling work, since the units measured are often smaller. Furthermore, the knowledge that had another sampling unit by chance been selected a very different value might have been obtained, may lead the inexperienced worker to believe that accuracy in the measurement of the selected units is of little importance.

When the sampling units are not natural units of the population, the selected units usually have to be demarcated at the time the measurements are taken. In crop sampling work in particular, where small areas are selected in order to obtain an estimate of the yield or other characteristics of the crop, location of the areas by means of randomly selected co-ordinates, though theoretically ensuring a random sample, will only in practice do so if the field work is carried out with complete objectivity. Since it is impossible in practice to locate the areas according to their co-ordinates by means of exact measurements, pacing or some similar approximate method must be used.

In this type of work the areas themselves should not be too small, both because errors in the demarcation of the boundaries become of increasing importance as the size of the unit is decreased, and also because the possibility of influencing the results by small changes in location, *e.g.* so as to include a particularly good plant, is greater the smaller the unit areas. Very small areas are capable of giving completely reliable results with experienced and well trained field-men, but may be very unreliable when used by inexperienced workers, particularly if the need for complete objectivity is not appreciated.

Sukhatme (1946a, H), for example, has reported the biases shown in Table 2.5 in some trial crop-sampling work on wheat. He himself expresses the opinion that the biases of the very small areas are due to the inclusion of border plants. This, however, would imply that the effective radius of the smallest areas, which were nominally circles of 2 ft. radius, would have to be increased by nearly 5 inches. Errors of this magnitude appear improbable,

unless the observers were very careless in their work, and it seems likely that part at least of the bias has been caused by faulty location.

Eye estimates are themselves a form of measurement, but such estimates are always subject to bias, which is likely to vary from observer to observer, and is often very substantial. If eye estimates are used, steps must therefore be taken to eliminate the resultant biases by carrying out proper measurements on a sub-sample of the material. A simple example of this is provided by the 1938-9 Census of Woodlands described in Section 4.25 and Examples 6.12.b and 7.11. A more complicated example is discussed in Sections 6.15 and 7.14.

TABLE 2.5—BIAS IN THE USE OF SMALL-SIZE AREAS IN SAMPLE SURVEYS FOR YIELD (*Sukhatme*)

Size of area in sq. ft.	No. of areas	Average yield in maunds per acre	Percentage overestimation
<i>Irrigated</i>			
471.5	78	10.10	—
117.9	78	10.58	4.8
29.5	78	11.69	15.7
28.3	117	11.60	14.9
12.6	117	14.38	42.4
<i>Unirrigated</i>			
471.5	107	6.55	—
117.9	107	7.27	11.0
29.5	107	8.08	23.4
28.3	162	7.52	14.8
12.6	161	9.33	42.4

2.6 Bias in estimation

In addition to biases which arise from faulty processes of selection and faulty work during the collection of the information, faulty methods of analysing the results may also introduce bias. A simple example occurs in the estimation of ratios. If, for instance, an agricultural crop is grown on types of land with different levels of fertility, and if the fields on the different types of land are of different average size, the mean yield per acre estimated from the mean of the yields per acre of all the fields may be markedly different from the mean yield per acre of all the land growing the crop. To take a numerical example, if there are three types of land having average yields of 20 cwt., 15 cwt. and 10 cwt. per acre respectively, and fields of an average size of 5 acres, 10 acres and 15 acres respectively, the number of fields on each

type of land being the same, the mean yield per acre over the whole of the land will be given by the *weighted* mean

$$\frac{5 \times 20 + 10 \times 15 + 15 \times 10}{5 + 10 + 15} = 13\frac{1}{3} \text{ cwt. per acre}$$

whereas the mean of the yields per acre of all fields will be 15 cwt. Consequently the bias in the estimate by the latter process will be about 12 per cent.

Biased estimates can be avoided by using the proper processes of estimation. This matter will be dealt with more fully in Chapter 6.

2.7 Circumstances in which bias is permissible

Although avoidance of any substantial bias is usually of the utmost importance, particularly in censuses on which administrative action has to be based, absence of bias is not always essential. In some types of investigation a certain amount of bias, provided it is reasonably constant, can be accepted. In censuses which are repeated at frequent intervals with a view to determining the changes rather than absolute values, for instance, a small overall bias may be of little consequence, provided it is constant in time. Similarly in surveys which have as their main objective the comparison of different groups of the population a bias which is approximately constant from group to group will be of little importance. The investigator must also avoid attaching exaggerated importance to minor sources of bias which, in fact, can only produce errors which are trivial relative to the random sampling error.

2.8 Methods of reducing the random sampling error

Once the absence of any important bias has been ensured, attention can be turned to the random sampling errors. These must clearly be sufficiently small to achieve the accuracy required.

Apart from errors due to bias, the simplest way of increasing the accuracy of the sample is to increase its size. Other things being equal, the random sampling error is approximately inversely proportional to the square root of the number of units included in the sample.

The accuracy attained will, however, depend not only on the number of units included in the sample, but also on the variability per unit; or, more strictly, on that part of the variability per unit which contributes to the sampling error. It is here that the complications of sampling procedure, both of design and of subsequent analysis, arise. By suitable processes of selection, which while imposing restrictions on fully random selection do not introduce bias into the results, the part of the variability per unit which contributes to the sampling error can often be substantially reduced, and the size of the sample required for a given accuracy thereby diminished.

The simplest type of restriction is that known as *stratification*. The population is "stratified" or divided into blocks of units in such a manner

that the units in each stratum or block are as similar as possible. Each of the strata is then sampled at random. If the same proportion is taken from each stratum, it is clear that each stratum will be represented in the correct proportion in the sample, and consequently differences between different strata are eliminated from the sampling error.

In addition to stratification there are a number of other devices, which will be discussed in more detail later, by which the accuracy of the sampling procedure can be increased, often very substantially. The three most important are: utilization of supplementary information, use of a variable sampling fraction (sometimes called "optimal allocation") and multi-stage sampling.

Utilization of supplementary information, that is information which is derived from sources other than the sampling scheme, or from a more extended sample than that on which information on the main characters is collected, takes a number of forms. A simple example will illustrate the general principle. Suppose that an estimate of the wheat yield of a country is required, and that a random sample of wheat fields has been taken and the total yield of each field determined. We can then estimate the total wheat yield of the country, either (a) by multiplying the total yield of the sample by the reciprocal of the proportion of the fields included in the sample, or (b) by calculating the mean yield per acre of the sampled fields (by dividing the total yield of all the sampled fields by their total area, so as to avoid bias) and multiplying this mean yield by the total acreage of wheat in the country. The latter estimate can only be made if the total acreage of wheat in the country is already known with sufficient accuracy, *e.g.* from returns made by the farmers or from a larger sample. If this information is available the second estimate is likely to be considerably more precise than is the first, since the variability of the total yields, which in so far as the yield per acre is constant will be proportional to the areas of the individual fields, is likely to be considerably greater than the variability of the yields per acre of the individual fields.

The use of a variable sampling fraction, *i.e.* the inclusion of different proportions of the different strata in the sample, enables the more important, or more variable, parts of the population to be sampled more intensively. If this is done it will of course be necessary to weight the contributions of the different strata to the total in the correct proportions.

The optimal sampling fractions depend on the relative variability of the different strata into which the population is divided for the purpose of taking a sample. Thus, if it is required to determine the number of workers in a given industry, it will be better to take a much larger fraction, possibly all, of the large factories than of the smaller factories.

In multi-stage sampling the population is divided into a number of first-stage sampling units, which are sampled in the ordinary manner, the selected first-stage units being subdivided into smaller second-stage units, which are also sampled. Further stages can be added if required. Thus, for example, in a population survey, a sample of all towns and villages may be taken, and in each of the selected towns and villages a sub-sample of all households may

be taken, with, possibly, for certain purposes, a further sub-sample of individuals from the selected households.

2.9 Choice of unit

In some classes of material there is considerable choice in the type and size of sampling units, and this gives further scope for increase in the efficiency of the sampling procedure. In general, when a given proportion of the material is included in the sample, the smaller the sampling units employed, the more accurate and representative will be the results. Thus, for example, in an agricultural survey, it will be more accurate to take 10 per cent. of all farms in each parish, or other small administrative unit, than to take all the farms in 10 per cent. of the parishes. This will remain true even if multi-stage sampling is adopted. It will be more accurate, for example, to take 10 per cent. of all the parishes in each county, with a second-stage sampling of the farms of each selected parish, rather than to take all the parishes in 10 per cent. of the counties, with the same degree of second-stage sampling. The reason for this is fairly obvious. The units in any region are likely to be more alike than are those of different regions, and if regions are used as sampling units, all the units in a region will be included or excluded from the sample simultaneously.

This need for small units distributed over the whole of the population often conflicts with the administrative requirements. It is clearly easier to arrange for a survey of farms in compact areas, such as parishes or counties, than to have to survey the same number of farms scattered over the whole country. The choice of a suitable balance between these two conflicting requirements is often one of the main problems in the planning of a sample survey. Furthermore, if only a small number of large units are included in the sample, whether or not there is second-stage sampling of these units, the sampling error will not be well-determined, since there will be relatively few differences between units on which to base the estimate of this.

We see, therefore, that the choice of sampling method depends not only on the relative accuracy of the different methods, but also on their practical convenience. It is important, for example, that the process of selection of the sample should not involve excessive preliminary work in the form of mapping, etc. The most suitable sampling method will therefore depend very much on the type of information that is already available on the population to be sampled; a method which may be excellent for a country where good maps are available may be entirely useless in a country which is inadequately mapped. Again, it is important that not only should the collection of the information not involve excessive travelling, but also it should be possible to subject the field-workers to proper supervision: consequently, sampling procedures which may be excellent with postal questionnaires may be entirely unsatisfactory when the information is collected by special investigators.

THE STRUCTURE OF VARIOUS TYPES OF SAMPLE

3.1 Definition of frame and sampling unit

In this chapter we propose to give a technical description of the structure of the various types of sample which are most commonly employed in practice, and the methods which must be followed in selecting them. The methods of obtaining estimates of the population values and of the sampling errors from the sample values will be discussed in Chapters 6 and 7.

All rigorous sampling demands a subdivision of the material to be sampled into units, termed *sampling units*, which form the basis of the actual sampling procedure. These units may be natural units of the material, such as individuals in a human population, or natural aggregates of such units, such as households, or they may be artificial units, such as rectangular areas on a map, bearing no relation to the natural subdivisions of the material.

It is not always necessary to make an actual subdivision of the whole of the material before selection of the sample, provided the selected units can be clearly and unambiguously defined. Thus, with sampling units which are rectangular areas on a map there is no need to demarcate all these areas; they can be defined by co-ordinates, and the selected areas demarcated after selection.

Clear and unambiguous definition demands the existence or construction of some form of *frame*. In the sampling of a human population, for instance, with households as sampling units, there must be available a list of all households, and this list must be such that any household selected from it can be unambiguously located. In area sampling from maps, the maps must be such that the selected areas can be unambiguously defined on the ground.

The specification of the frame implicitly defines the geographical scope of the survey and the categories of material covered. A survey of a human population based on a list of households, for instance, will only cover those categories of the population which constitute the households included in the list. If other categories require inclusion, or if the frame is defective, special steps will have to be taken to supplement and emend it.

In statistical terminology any aggregate of values is termed a *population*, and consequently the whole aggregate of sampling units into which the material is divided is known as the *population of sampling units*. If the sampling units are aggregates of the natural units of the material, these natural units will form a further population which must be distinguished from the population of sampling units.

In America the term *cluster sampling* has been applied to sampling in which the sampling units are aggregates or "clusters" of the natural units. The term is a somewhat loose one, since there is often a hierarchy of natural units, e.g. a sample in which the sampling units are households may be regarded as an ordinary sample of households or as a cluster sample of individuals.

In multi-stage sampling there is also a hierarchy of sampling units, first-stage, second-stage, etc., corresponding to the different sampling stages, and each set of units will form its own population of units.

Sampling units may be of the same or differing size. They may contain the same, or approximately the same, number of natural units, or they may contain widely differing numbers. The whole procedure of sampling, including the estimation of the population values and the sampling errors, is simplest when the sampling units are of approximately the same size and contain approximately the same number of natural units. Often, however, the material is such that this condition cannot be conveniently fulfilled. In particular, if the natural units are themselves of widely differing size, variation in size of the sampling units or in the number of natural units they contain is inevitable.

There is nothing in the sampling process which demands that the sampling units should be of any particular size, but, as has been explained in Section 2.9, the smaller the sampling units employed the more accurate will be the results obtained when a given proportion of the material is included in the sample.

3.2 Random sample

A random sample is the simplest type of rigorously selected sample, and is the basis of most of the more complicated sampling methods. In a random sample, after subdivision of the material into sampling units, the requisite number of units are selected at random from the whole population of units.

As has been emphasized in Section 2.3, random selection implies a strict process of selection equivalent to that of drawing lots. In practice it may be carried out either by some such process, or preferably, since adequate shuffling of cards, etc., is difficult, by the use of a table of random numbers. A small table of random numbers is given at the end of the book. The examples of this section illustrate the use of such a table.

The process of random selection may proceed in two stages. Suppose that the population is divided into groups of units containing $x_1, x_2, x_3, \dots, x_n$ units. The successive sub-totals

$x_1 = X_1, x_1 + x_2 = X_2, x_1 + x_2 + x_3 = X_3, \dots, x_1 + x_2 + \dots + x_n = X_n$ are first calculated, which is easily done on a printing adding machine. The requisite number of numbers are then selected at random between 1 and X_n , numbers that occur more than once being rejected. A selected number that is greater than X_{s-1} , but less than or equal to X_s , indicates that a unit of the s th group is to be taken. Selection of a unit at random from this group, which can if convenient be made on the basis of the number already selected, will then give the equivalent of completely random selection.

This two-stage process is of value when the full numbering or demarcation of the units in all groups before sampling is laborious, since only the total number of units in each group need be known. It is of particular value when the units are artificially demarcated areas, and the total areas of natural

subdivisions of the material are known. By using the process only the units in the selected groups have to be numbered or demarcated.

Example 3.2.a

Select a sample of 20 from a population of 2879 units.

Using the four-figure numbers given by the first four columns of digits in Table A. 1, and rejecting all numbers greater than 2879, we obtain the sample 347, 1676, 1256, 1622, 1818, 2662, 2342, 1608, 2742, 39, 1690, 1127, 1490, 2046, 526, 797, 2699, 1465, 2467, 1753.

The above procedure results in the rejection, in this example, of nearly three-quarters of the random numbers given by the table. Various devices may be used to avoid this. In the present example the simplest is to take the numbers 3001-6000 and 6001-9000 as equivalent to 0001-3000, rejecting the numbers 9001-9999 and 0000. Using the second column of four-figure numbers gives the sample 1373, 2467, 227, 2599, 2635, 1794, 1753, 378, 1234, 2632, 792, 897, 1064, 2819, 1712, 1837, 2722, 1504, 13, 2565.

If with either of the above procedures the same unit is selected a second time, the number leading to this selection is rejected, and an additional number taken.

It will be noted that neither of these samples is evenly distributed over the whole range of units. The distribution between the different thirds of the range is in fact :

Numbers	1st sample	2nd sample
1-960	4	5
961-1920	10	8
1921-2879	6	7
	<u>20</u>	<u>20</u>

Random selection will give samples that deviate somewhat from an even distribution, the actual deviations being themselves governed by statistical laws. Exact statistical tests show that about three out of four samples will have smaller aggregate deviations than the first sample, but only three out of ten will have smaller aggregate deviations than the second sample.*

Example 3.2.b

Select unit areas $\frac{1}{10}$ mile $\times$ $\frac{1}{10}$ mile at random from a rectangular area 5 miles $\times$ 4 miles.

There are 2000 unit areas, which can best be defined by co-ordinates 1-50 along the longer side of the rectangle, and 1-40 along the shorter side, the

* The appropriate test is that known as the χ^2 test. A description of this test will be found in most modern statistical textbooks.

co-ordinates selected defining the corner of the unit area furthest from the corner of the rectangle (0, 0). The selection of a number at random between 1 and 50, and a second between 1 and 40, will therefore select a unit at random. Taking the third column of four-figure numbers (beginning 8636) and following the second of the procedures of Example 3.2.a gives the pairs of co-ordinates 36, 36 ; 12, 02 ; 16, 16 ; 14, 38 ; etc.

If points instead of unit areas are to be selected each co-ordinate range should theoretically be infinitely subdivided. The actual degree of subdivision need not usually be very fine.

The procedure of this example may be used for the selection of unit areas or points from an irregularly shaped area, provided the extreme range of each co-ordinate is included, points falling outside the area being rejected. More elaborate processes, involving less rejection, can of course be devised, but care must be taken that the probability of selection of all areas or points is equal. Thus in a triangular area, the selection of lines parallel to the base at random distances from the base, followed by the selection at random of a point within the triangle on each of the selected lines, will give a greater density of points near the apex of the triangle. The selection of points within a circle by the selection of random distances and bearings from the centre will give a greater density of points near the centre. In irregularly shaped areas, also, fractional unit areas requiring special treatment will occur at the boundaries.

Example 3.2.c

14 streets in a ward contain 25, 17, 5, 59, 64, 22, 38, 16, 21, 12, 14, 38, 17, 23 houses respectively. Make a random selection of 6 houses from all 371 houses.

The successive sub-totals are 25, 42, 47, 106, 170, 192, 230, 246, 267, 279, 293, 331, 348, 371. A table of random numbers gives the numbers 72, 128, 96, 326, 199, 202. The units 72 and 96 therefore fall in the 4th street, the unit 128 in the 5th street, the units 199 and 202 in the 7th street, and the unit 326 in the 12th street. Since $72 - 47 = 25$, and $96 - 47 = 49$, the 25th and 49th houses in the 4th street are selected, etc. The numbering of four streets, involving 199 houses, is required.

3.3 Stratification with uniform sampling fraction

In a stratified sample the population of sample units is subdivided into groups or "strata" before selection of the sample. These strata may all contain the same number of units, or differing numbers of units. If a uniform sampling fraction is used, the same fraction of the units of each stratum is included in the sample, the units selected being chosen at random from all the units within each stratum. A stratified sample is thus equivalent to a set

of random samples on a number of sub-populations, each equivalent to one stratum.

Stratification has two purposes. The first is to increase the accuracy of the overall population estimates. The second is to ensure that subdivisions of the population which are themselves of interest are adequately represented. Such subdivisions may be termed domains of study. Maximum overall accuracy will be attained if the strata are so chosen that the units within each stratum are as similar as possible. It will often be advisable to use domains of study as strata, however, even if some other form of stratification might be expected to give somewhat more accurate results. If there is marked heterogeneity within some or all of the domains of study these may themselves be subdivided into smaller strata for the purposes of sampling.

Stratification affects the estimation of the sampling error. Since in a stratified sample only variation within strata gives rise to sampling error, it is this component of variation that requires estimation, and this can in general only be done from differences between units in the same stratum. It is therefore necessary, if an estimate of sampling error is required, that the strata be of such size that the sample contains two or more units from at least the majority of strata. In certain cases, in which the use of strata containing only a single selected unit appears advisable on account of the gain in accuracy thereby obtained, special methods, often of an approximate nature, of estimating the sampling error have to be adopted. The matter is discussed further in Section 8.15.

If the sampling units are already classified in the required strata, the selection of a stratified sample can be made in the same way as a random sample, the requisite number of units being selected at random from each stratum. If, however, the population is not so classified, selection by this method would necessitate prior classification. In this case, if the numbers of units in the different strata are known, an alternative procedure is available. This consists of selecting a sample at random; keeping a tally, as the selection proceeds, of the numbers falling in each stratum; and rejecting any further members of a particular stratum as soon as the requisite number for that stratum has been obtained. On the other hand, if the numbers of units in the different strata are not known, a count covering the whole population will in any case have to be made, in which case a classification which will serve as a basis for the subsequent selection of the sample may well be carried out simultaneously.

Unless all the strata contain the same number of units it will usually happen that the chosen sampling fraction will not give an exact whole number of units in each stratum. In this case the nearest whole number of units has to be taken. We may thus differentiate between the working sampling fraction, which with stratification with a uniform sampling fraction is the same for all strata, and the exact sampling fractions, which will differ slightly from the working sampling fraction. The use of the working sampling fraction in the analysis of the results leads to minor inaccuracies, but these will seldom give rise to errors of any practical importance.

It may be noted that if the numbers of units from the whole population falling in different strata are known, and a random sample is taken which is sufficiently large to ensure that adequate numbers of units are obtained from all strata, adjustment of the results so that the different strata are represented in their correct proportions will lead to practically the same accuracy as would be obtained with a stratified sample. Which of these two alternative courses is adopted in any particular case is a question of convenience. If the selection of either type of sample is equally simple it is best to use a stratified sample, as the computations are thereby simplified. In certain cases, however, the classification of units into strata may only be possible by means of information obtained in the course of the survey, in which case a random sample, with subsequent adjustment, is required. Thus, for example, in a survey of a human population, the age distribution of the whole population may be known, but prior selection of individuals of particular ages may be impossible owing to lack of information on these ages.

3.4 Multiple stratification

A population may be stratified for two or more different characteristics. If selection is made from sub-strata formed of the various combinations of the main classifications the procedure is exactly equivalent to ordinary stratification, the sub-strata being equivalent to strata. Thus we may stratify farms according to size and according to geographical regions. If the farms in each region are classified into size-groups before taking the sample then the region-size-group combinations form the individual sub-strata.

Occasionally the number of units of the population falling in each set of main strata may be known, *e.g.* from prior census data, but not the numbers in the various sub-strata. Thus, in the above example there may be information on the numbers of farms in the different size-groups, and also on the numbers in the different geographical regions, but not on the numbers of each size-group in each region. In such cases we may attempt the selection of a sample which will have the right proportions for each set of main strata. Such stratification may be termed *multiple stratification without control of sub-strata*. The selection of such a sample, however, presents both theoretical and practical difficulties, and the calculation of the sampling error is also troublesome.

In the rare cases in which multiple stratification without control of sub-strata is deemed to be necessary a simple procedure of selection which should give a reasonably satisfactory sample is as follows. Units are selected at random until the total of every row and column of the two or more-way table for the sets of strata is at least equal to the required total. The excesses of these marginal totals are calculated, and numbers chosen for deduction from the sub-strata totals which together make up these excesses, and which, subject to these restrictions, are about proportional to the sub-strata totals. (A method of calculating such numbers is shown in the following example.) The corresponding numbers of units are then rejected from the sub-strata groups,

those rejected being selected at random. If the original selection was strictly random the condition of randomness will be fulfilled if those last selected are rejected.

Example 3.4

A sample of 1000 is required from a population classified into two sets of four strata, the sub-strata totals being unknown, but the correct strata totals for the sample being known to be 120, 280, 350, 250, for each set of strata.

After a sample of 1125 units had been drawn, the numbers of units in the 16 sub-strata shown in Table 3.4.a were obtained.

TABLE 3.4.a—TWO-WAY STRATIFICATION WITHOUT CONTROL OF
SUB-STRATA : INITIAL SAMPLE

Strata B	Strata A				Total	Required	Excess
	1	2	3	4			
1	37	40	35	8	120	120	0
2	39	140	82	56	317	280	+ 37
3	45	97	173	93	408	350	+ 58
4	8	40	86	146	280	250	+ 30
Total	129	317	376	303	1125	1000	125
Required	120	280	350	250	1000	—	—
Excess	+ 9	+ 37	+ 26	+ 53	125	—	—

TABLE 3.4.b—CALCULATION OF NUMBERS OF UNITS TO BE REJECTED

	Stage 1					Stage 2				
	A_1	A_2	A_3	A_4	Total	A_1	A_2	A_3	A_4	Total
B_1	0	0	0	0	0	0	0	0	0	0
B_2	5	16	9	7	37	- 1	+ 2	- 4	+ 3	0
B_3	6	14	25	13	58	- 2	+ 1	- 9	+ 5	- 5
B_4	1	4	9	16	30	0	0	- 4	+ 9	+ 5
Total	12	34	43	36	125	- 3	+ 3	- 17	+ 17	0
Required	9	37	26	53	125					
Difference	+ 3	- 3	+ 17	- 17	0					

TABLE 3.4.b—Continued

	Stage 3				
	A_1	A_2	A_3	A_4	Total
B_1	0	0	0	0	0
B_2	0	0	0	0	0
B_3	+1 (0)	+1	+2	+1 (+2)	+5
B_4	0	-1	-1 (-2)	-3 (-2)	-5
Total .	+1 (0)	0	+1 (0)	-2 (0)	0

TABLE 3.4.c—NUMBERS OF UNITS REJECTED, NUMBERS IN FINAL SAMPLE AND CORRECT SUB-STRATA TOTALS

	Numbers of units rejected					Numbers in final sample				
	A_1	A_2	A_3	A_4	Total	A_1	A_2	A_3	A_4	Total
B_1	0	0	0	0	0	37	40	35	8	120
B_2	4	18	5	10	37	35	122	77	46	280
B_3	4	16	18	20	58	41	81	155	73	350
B_4	1	3	3	23	30	7	37	83	123	250
Total .	9	37	26	53	125	120	280	350	250	1000

	Correct sub-strata totals				
	A_1	A_2	A_3	A_4	Total
B_1	40	40	30	10	120
B_2	40	120	80	40	280
B_3	30	80	160	80	350
B_4	10	40	80	120	250
Total .	120	280	350	250	1000

The three stages of the calculation are shown in Table 3.4.b. In stage 1 the excesses of the rows have been distributed in proportion to the numbers of units in the sub-strata of each row. The distributed excesses are added by columns and compared with the required excesses. The differences, with signs reversed, are distributed by columns in stage 2, excluding the first row,

and the process is repeated for rows in stage 3. The numbers are now small and empirical adjustments, shown in brackets, have been chosen to make the column totals zero.

The three stages are then summed and deducted from the numbers in the original sample, as shown in Table 3.4.c. This table also shows the correct proportional sub-strata totals of the population from which the sample was drawn. The appropriate statistical test* shows that a satisfactory sample has been obtained.

The above process does not necessarily converge, but will usually do so in practical cases. If a negative value for number of units rejected is obtained for any sub-stratum this can be brought to zero by an empirical adjustment.

3.5 Stratification with a variable sampling fraction

In certain types of material very considerable gains in accuracy will result if different sampling fractions are used for the different strata. The greatest accuracy for a given number of units will be attained if the sampling fractions are proportional to the within-strata standard deviations† of the units. If the sampling fractions are denoted by $f_1, f_2, \dots$ and the standard deviations by $\sigma_1, \sigma_2, \dots$ we have

$$\frac{f_1}{\sigma_1} = \frac{f_2}{\sigma_2} = \dots$$

In some cases this formula may give sampling fractions greater than unity for some of the strata. If this occurs the whole of these strata are included in the sample.

A particularly important application of the variable sampling fraction is to material stratified into size-groups. In such material the various quantitative characteristics of the units under investigation often have within-strata standard deviations which are roughly proportional to the mean sizes of the units in the different size-groups. In this case the sampling fractions should be taken about proportional to these mean sizes. If quantitative characteristics very highly correlated with size of unit are under investigation, the ranges of the size-groups may give good estimates of the relative within-strata standard deviations. The sampling fractions may then be taken proportional to these ranges. Changes with time, however, are usually by no means so highly correlated with size, and when the changes are of interest, sampling fractions proportional to the mean sizes of the size-groups will usually be best.

The above rules will determine sampling fractions which give the maximum accuracy for estimates of the population values. In cases in which the values for the individual strata are of interest, *i.e.* cases in which the strata themselves form domains of study, it is also important to see that all the strata are adequately

* $\chi^2 = 9.1$, 9 degrees of freedom.

† The meaning of this term and the method of estimation will be explained in Chapter 7.

represented in the sample, and for this reason the rule of strict proportionality to the standard deviations, or to the mean sizes of the size-groups, often requires some modification.

When several quantities are under investigation, it will usually be found that their within-strata standard deviations are not in quite the same proportions. This, however, is not a very serious problem in practice, since any sampling fractions which are somewhere near the optimal will give results which are nearly as accurate as those given by the optimal fractions. Consequently there is usually no great difficulty in choosing suitable sampling fractions which will reconcile the various conflicting requirements.

Since, for these reasons, sampling fractions are often used which are not optimal, we have preferred not to adopt the term "optimal allocation," which has sometimes been used to denote stratified sampling with a variable sampling fraction.

The within-strata standard deviations can only be estimated from data relating to the material to be sampled, or from data derived from similar material, but general knowledge of the behaviour of material of a particular type, *e.g.* material stratified into size-groups, will often enable suitable sampling fractions to be chosen with all necessary accuracy. It is sometimes suggested that a preliminary survey should be undertaken merely to determine the optimal sampling fractions, but this is rarely worth while, though if a preliminary survey is being undertaken for other purposes it will of course also serve to improve the sampling fractions.

3.6 Systematic samples from lists

Although the importance of the principle of random selection in sampling has been stressed, much practical sampling is in fact not fully random in character. Thus a frequent method of selecting a sample, when a list of the units of the population to be sampled is available, is to take every q th entry on this list. This may be termed a systematic sample from the list. Other more complicated systematic procedures may occasionally be adopted for special purposes.

It is customary, and salutary, to determine the first entry by selecting a number at random between 1 and q , but this element of randomness does not convert the sample into a random one.* A systematic sample would be equivalent to a fully random sample if the list were arranged wholly at random. No lists, however, are arranged at random. The nearest approach to random order is probably provided by alphabetical lists, though even these have certain non-random characteristics: in this country, for instance, a large proportion of the Scotsmen will be found under the letter M. If every q th entry is taken, a kind of partial stratification will therefore be obtained, and the sample will be somewhat more precise than a fully random sample. Thus in a

* Except in the trivial sense that the sample is a random sample of 1 unit out of q units, each unit being composed of the aggregate of a set of all entries at spacing q .

systematic sample of farms taken from a list of farms arranged by parishes, the proportion of farms drawn from each parish will be more or less constant, provided the sampling interval is small compared with the number of farms in a parish.

Owing to the lack of definition of the strata it is impossible to make a fully valid estimate of the sampling error, but provided there are no periodic features in the list the sample will not be biased. An estimate of the sampling error which is good enough for practical purposes can usually be made by regarding the sample as a sample stratified in the major subdivisions of the list, ignoring any minor and ill-defined groupings. If the sampling error is estimated as if the sample were fully random an overestimate will be obtained, the inaccuracy being greater the more marked the similarity of the neighbouring entries in the list.

In general, systematic sampling from lists will be found to be quite satisfactory provided care is taken to see that there are no periodic features in the list which are associated with the sampling interval. The method is often much more convenient than random or stratified random sampling, since the labour of making a proper random selection, which in an extensive sampling scheme is often very considerable, is avoided. It must be clearly recognized, however, that the responsibility for the judgment that the material is such that systematic sampling will give satisfactory results rests with the investigator.

Sampling in which the selection is wholly systematic should be clearly distinguished from sampling in which there is proper random selection of sampling units which are themselves systematic aggregates of smaller subdivisions of the material. Thus a common method of sampling rows of potatoes has been to use sampling units consisting of every 20th plant, two such sampling units being selected at random from each row by selecting two numbers at random between 1 and 20. Such a method of sampling fulfils all the conditions required for fully valid random sampling.

3.7 An example of alternative ways of sampling highly variable material

In order to illustrate the various methods of sampling which have been discussed in the preceding sections, we will consider their application to the problem of determining the area under wheat in an English county. For this purpose the wheat acreages of Hertfordshire farms for 1939 were utilized.

The choice of problem and of basic material were dictated partly by convenience, and partly by the need for a complete set of data covering highly variable material, since an investigation of different sampling methods can be carried out most easily when data are available for the whole of the material.

It should be emphasized that this example is to be regarded as illustrative only. Sampling methods are very unlikely to be required for the determination

of crop acreages in a country such as England, since all farmers make returns of their acreages each year ; the only possible use of sampling would be at the compilation stage, where it might be used to avoid the necessity of totalling the whole of the returns. The present investigation is not fully relevant to this problem, however, since the acreage of only one of the most extensively grown crops is considered, and that for only one county. Considerably greater errors, proportionately, may be expected in the less common crops.

Records were available for 2496 farms, and the acreage of wheat, and also the total acreage of crops and grass, which is virtually the total acreage of farmed land and will be termed the *size* of the farm, were abstracted for each farm. The original records were arranged by districts, by parishes alphabetically within districts, and by farmer's name alphabetically within parishes. This order was preserved in the abstract. The return for any farm, or "holding," does not necessarily relate solely to land in a given parish, but may include land in other parishes farmed by the same farmer ; farmers with two or more distinct farms may make separate returns for these farms or may include them all in a single return. The total area of wheat in the county, from the abstracted returns, was 44,676 acres, and the total area of crops and grass was 273,074 acres.*

If farms are taken as sampling units the dominant source of variation in wheat acreage will be variation in size of farm, since farms range from 1 acre to over 1000 acres, and no farm can have more than a fraction of its area under wheat. Stratification by size of farm is therefore indicated. The use of a variable sampling fraction will also be advantageous, since the wheat acreages of the large farms will be much more variable than those of the smaller farms. Further stratification by districts is possible, but is not likely to give much increase in precision unless the incidence of wheat growing in the different districts is very markedly different. In any case a systematic method of selection from the list, which in view of the alphabetical method of arrangement will be quite satisfactory, will give the effect of stratification by districts.

For comparative purposes the following samples were taken :

- (1) a random sample of 1 in 20 farms, 125 farms in all ;
- (2) a stratified random sample with a uniform sampling fraction of 1 in 20 ;
- (3) a stratified systematic sample with a variable sampling fraction, the fraction being approximately proportional to mean size of farm within each size-group, and chosen so as to give about the same number of farms, actually 135, as samples 1 and 2 ; the systematic method of selection within size-groups results in approximate stratification by districts also.

* It may be noted that these values disagree with the values shown in *Agricultural Statistics*, viz. 46,281 acres and 278,380 acres respectively. The reasons for this discrepancy need not concern us here, but it provides an illustration of the fact that disagreement between sample and complete returns must not be assumed to be necessarily solely due to sampling error.

The size-groups chosen, the number of farms in each group, and the sampling fractions and numbers of farms for sample 3 are shown in Table 3.7.a. For sample 2 the last two size-groups were combined.

TABLE 3.7.a—HERTFORDSHIRE FARMS, 1939: SIZE-GROUPS, NUMBERS OF FARMS, AND CHOSEN VALUES OF THE VARIABLE SAMPLING FRACTION

Size-group (acres crops and grass)	No. of farms in county	Sampling fraction	No. of farms in sample
1-5	435	Nil	0
6-20	519	1/200	3
21-50	357	1/60	6
51-150	519	1/20	26
151-300	400	1/10	40
301-500	215	1/5	43
501-	51	1/3	17
	2,496		135

As pointed out in Section 3.3, the random sample can be stratified after selection so as to eliminate the effect of variation between size-groups, provided the number in each size-group of the population is known. Variation due to size may also be eliminated by using size as supplementary information, provided information on size of farm is available for each farm in the sample, and the total area of all farms in the county is known. Either the ratio or the regression method may be used, as explained in Chapter 6.

The estimates of the wheat acreage and of the number of farms growing wheat obtained from these samples, their estimated sampling standard errors, and in the case of the acreage the actual errors of the estimates, are shown in Table 3.7.b. The sampling standard errors, as will be explained in detail in Chapter 7, give measures of the average magnitudes of the sampling errors that may be expected with given methods of sampling and of estimation. The computations leading to these estimates are set out in Chapters 6 and 7, where tables giving the actual values of the sample units for the three samples (Tables 6.6.a, 6.5.a and 6.7.a respectively) will also be found.

It is apparent from the values of the sampling standard errors for wheat acreage that, as is to be expected, both stratification and the use of a variable sampling fraction have resulted in large gains in accuracy.* The use of supplementary information on size, whether by stratification after sampling, by ratio or by regression, serves to make the random sample about as accurate

* The word *accuracy* is used in this book to denote the expected accuracy of an estimate, as indicated by its sampling standard error. It is occasionally also used to denote the actual accuracy, as indicated by the actual error (usually unknown), but this should not cause any confusion.

as the stratified sample with a fixed sampling fraction. This is also to be expected.

The numbers of units, *i.e.* farms, required to attain the same accuracy with different methods of sampling and estimation may be taken as roughly proportional to the squares of the standard errors, allowance being made for the greater number of units included in sample 3. These are shown in the column "relative variance," the random sample with direct estimation

TABLE 3.7.b—HERTFORDSHIRE FARMS: COMPARISON OF VARIOUS TYPES OF SAMPLES AND METHODS OF ESTIMATION

No.	Method of selection and estimation	Wheat acreage				No. of farms growing wheat		
		Estimate	Standard error	Actual error	Relative variance per farm	Estimate	Standard error	Relative variance per farm
1a	Random, direct	46,020	$\pm 7,950$	+ 1,340	100	900	± 104.6	100
1b	Random, stratified after selection	41,100	$\pm 4,320$	- 3,580	30	860	± 75.2	52
1c	Random, by ratio	41,570	$\pm 3,940$	- 3,110	25	Not calculated		
1d	Random, by regression	40,400	$\pm 4,130$	- 4,280	27	Not calculated		
2	Stratified, direct	40,220	$\pm 4,110$	- 4,460	27	1,080	± 71.6	47
3	Variable sampling fraction, direct	42,765	$\pm 2,550$	- 1,915	10	911	± 88.9	72

being set at 100. Thus stratification by size, or elimination of variation due to size in the estimation process, reduces the number of farms required by a factor of about 4, and the variable sampling fraction results in a further reduction by a factor of about $2\frac{1}{2}$.

The situation with respect to number of farms growing wheat is somewhat different. Stratification has again resulted in considerable increase in accuracy, though the gain is not so great as with acreage. The sample with a variable sampling fraction, on the other hand, is not so accurate as the ordinary stratified sample. The sampling fractions which are optimal for the determination of

wheat acreage are by no means optimal for the determination of the number of farms growing wheat.

The actual errors of the estimates bear little relation to the sampling standard errors, except that they are in no case markedly larger than these standard errors. The random sample without adjustment gives the most accurate estimate of acreage, though this estimate has the largest sampling standard error, and adjustment of the random sample makes the actual error larger, though it reduces the sampling standard error. This is an illustration of the fact that an inaccurate method of sampling will sometimes by chance give an accurate estimate. The accuracy of a sampling procedure must never be judged by the magnitude of a single discrepancy; a large discrepancy provides some evidence that a method is inaccurate, but a single small discrepancy provides practically no evidence that it is accurate.

3.8 Multi-stage sampling

In multi-stage sampling the material is regarded as made up of a number of first-stage sampling units, each of which is made up of a number of second-stage units, etc. The sampling process is carried out in stages. At the first stage the first-stage units are sampled by some suitable method, such as random or stratified sampling. At the second stage a sample of second-stage units is selected from each of the selected first-stage units, again by some suitable method, which may be the same as or different from the method employed for the first-stage units. Further stages may be added as required.

Multi-stage sampling introduces a flexibility into sampling which is lacking in the simpler methods. It enables existing natural divisions and subdivisions of the material to be utilized as units at the various stages, and, as pointed out in Section 2.9, it permits the concentration of the field work of censuses and surveys covering large areas. On the other hand, for the reasons there given, a multi-stage sample is in general less accurate than is a sample containing the same number of final-stage units which have been selected by some suitable single-stage process.

Multi-stage sampling also has the important advantage that subdivision into second-stage units, *i.e.* the construction of the second-stage frame, need only be carried out for those first-stage units which are actually included in the sample. It is therefore particularly valuable in surveys of undeveloped areas where no frame exists which is sufficiently detailed and accurate for subdivision of the material into reasonably small sampling units.

Since there are many variants of multi-stage sampling which are possible for any given type of material, careful investigation is often required before a decision as to the procedure which is best for any particular purpose can be reached. This matter will be discussed in detail in Chapter 8, after the methods of evaluating sampling errors have been described.

3.9 Sampling with probabilities proportional to size of unit

If we have areas demarcated on a map, such as fields, and a point is located at random on the map, the probabilities of the point falling within the boundaries of the different fields are clearly proportional to the areas of the fields. Consequently areas can be selected at random with probabilities proportional to their size by the simple procedure of taking random points on the map. It will be noted that such a process of selection may result in the same area being included twice or more in the sample. In this case it must be counted twice or more. We cannot, without distorting the probabilities, make a further selection in the manner followed with equal probabilities.

The principle has applications in agricultural surveys designed to determine the acreage and yield of different agricultural crops, total cultivated area, etc. All that is required for acreage is to determine the proportion of points which fall in areas of the given type. The method is therefore particularly attractive when carrying out surveys of the areas of crops, etc., by aerial survey, provided the different crops can be recognized on the photographs, since it avoids all the measurements of area which would be required if an ordinary random sample of areas were taken. The sampling of the fields with probabilities proportional to size is in this case equivalent to the sampling of small unit areas of equal size whose locations are determined by the random points. When only areas require to be determined the sizes of the fields in which the random points fall are in fact immaterial.

The analogy with the case of a stratified sample with a variable sampling fraction indicates that under certain circumstances greater precision may be expected from areas selected with probabilities proportional to size than will be obtained if they are selected with equal probabilities.

In the case of yield determinations, when the total acreage is known, the determinations of the yield from a sample of fields selected with probability proportional to size may always be expected to give a more accurate estimate of the mean yield per acre and total yield than will similar yield determinations on a random sample of fields irrespective of size. If the total acreage is not known then the situation is more complicated, but here again sampling with probabilities proportional to size is often advantageous.

Sampling with probabilities proportional to size of unit, or to some other known quantitative character of the units, may be carried out on other types of material by forming a cumulative or running total of the sizes of the units, and selecting numbers at random from the total of all the units in the manner of Example 3.2.c. Stratification by size and the use of a variable sampling fraction will usually be preferable in such cases, however, on the grounds both of accuracy and convenience, except in the special circumstances to be described in the next section.

3.10 Sampling from within strata with probabilities proportional to size of unit

Apart from area sampling, sampling with probabilities proportional to size of unit is mainly of use when the units are stratified according to some other characteristic, and the number of units to be selected from each stratum, or from some of them, is small. In this case an ordinary stratified sample will give either inaccurate or biased estimates when the ratio method of estimation, explained in Chapter 6, is used. The bias or inaccuracy is removed by selecting the units from within strata with probabilities proportional to size. This fact appears to have been first recognized by Hansen and Hurwitz (1943, A).

This procedure is particularly useful in conjunction with two-stage sampling with large first-stage units of variable size. The first-stage units are selected from within strata with probabilities proportional to size, and the second-stage units are selected with probabilities inversely proportional to the sizes of the first-stage units. By this device the overall sampling fraction is kept constant, with consequent simplification of the computations.

As before, when more than one unit is required from a stratum, selection with probability strictly proportional to size can only be simply effected if a unit which happens to be selected twice is counted twice. Generally, however, when each stratum contains only a few large units, duplication of units is not desirable; instead a further unit is selected, with slight inexactitude in the probability of selection, and consequent slight, but usually negligible, bias in the results.

3.11 An example of sampling by administrative areas

Reverting to the illustrative example of Hertfordshire farms considered in Section 3.7, we may now investigate the effect of taking parishes as sampling units, or as first-stage units in two-stage sampling with farms as second-stage units.

The use of parishes as sampling units may be expected to result in a sample which is somewhat less accurate than a sample containing the same number of farms distributed over all parishes. Nevertheless, if actual visits have to be made to the farms it may pay to use parishes as sampling units, or as first-stage units in two-stage sampling. In analogous situations in undeveloped countries, where definition of farm boundaries may present difficulties, the complete survey of small administrative or other areas may also be better than any attempt to sample individual farms.

Inspection of the Hertfordshire data showed that the total farm area (crops and grass) included in the returns of farmers of a single parish was very variable, partly owing to variations in size of the parishes, and partly because some of the parishes were mainly urban in character. It is therefore best to use individual parishes as sampling units only if they contain a certain minimum acreage of crops and grass. The minimum chosen was 2000 acres, the remaining

parishes being grouped roughly in the order in which they appeared in the alphabetical list, to form "combined" parishes containing over the minimum acreage. The effect of this combination is shown in Table 3.11.a.

TABLE 3.11.a—NUMBERS OF FARMS, PARISHES, AND "COMBINED" PARISHES
IN THE DISTRICTS OF HERTFORDSHIRE

District No.	District	No. of farms	No. of parishes	No. of parishes after combination
1	Barnet	230	17	7
2	Bishop's Stortford .	316	23	16
3	East Herts. . . .	564	31	20
4	Hitchin	553	36	25
5	St. Albans	218	9	7
6	Tring	424	16	11
7	Watford	191	8	5
		2,496	140	91

Districts were used as strata in this sampling, 1 in 5 "combined" parishes being taken per district, *i.e.* 17 parishes in all.

Two samples were taken. In sample *A* the parishes were selected in the ordinary manner, with equal probability of selection for each parish. In sample *B* selection with probabilities proportional to size was employed. The parishes of sample *B* were also sub-sampled in two ways, samples B_1 and B_2 . In sample B_1 a uniform sampling fraction of $\frac{1}{4}$ was taken for sampling at the second stage, with stratification by size, using the size-groups of Table 3.7.a with the last two size-groups combined. In sample B_2 a variable sampling fraction was used with values $\frac{1}{10}$ for size-group 1–50 acres, $\frac{1}{4}$ for 51–150 acres, $\frac{1}{2}$ for 151–300 acres, and 1 for over 300 acres. Sample *B* is given in detail in Example 6.17.

The relative efficiency of the various methods is discussed in Section 8.9. The results are summarized in Table 3.11.b. This table is similar to Table 3.7.b, except that estimated average values of the standard errors are given, and not those calculated from the actual selected samples. These latter are not sufficiently accurate for comparison owing to the small number of parishes involved.

It will be seen that a sample of 1 in 5 parishes provides results which are decidedly more accurate than a stratified random sample of 1 in 20 farms with a uniform sampling fraction, but somewhat less accurate than a similar sample with a variable sampling fraction. The stratified random sample of 1 in 20 farms is 1.29 times as accurate as sample B_1 , allowing for differing numbers of farms. The similar sample with the variable sampling fraction is 1.83 times as accurate as sample B_2 . Sample *B* is somewhat more accurate

than sample *A*, particularly if the unbiased estimate given by the overall ratio is used for sample *A*. The difference is not marked, however, since the combination of parishes has created units which do not differ excessively in size.

TABLE 3.11.b—HERTFORDSHIRE FARMS: SAMPLES FOR WHEAT ACREAGE WITH "COMBINED" PARISHES AS SAMPLING UNITS OR FIRST-STAGE UNITS IN TWO-STAGE SAMPLING

Sample	No. of stages	Method of Sampling		Method of estimation	Estimate	Expected standard error	Actual error	Relative variance
		1st stage	2nd stage					
<i>A</i>	1	Stratified by district	—	Overall ratio	41,730	± 3,080	— 2,950	100
				District ratios	41,010	± 3,010	— 3,670	95
<i>B</i>	1	Stratified by district, probability proportional to size	—	District ratios	46,660	± 2,870	+ 1,980	87
<i>B</i> ₁	2	„	Stratified random	District ratios	48,930	± 4,950	+ 4,250	259
<i>B</i> ₂	2	„	Variable sampling fraction	District ratios	45,600	± 3,460	+ 920	127

3.12 Multi-phase sampling

It is sometimes convenient and economical to collect certain items of information from the whole of the units of a sample and other items of information from some only of these units, these latter units being so chosen as to constitute a sub-sample of the units of the original sample. This may be termed *two-phase sampling*. Further phases may be added if required.

Multi-phase sampling is of value in several ways. Its simplest application is to the case in which the number of units needed to give the required accuracy on different items is widely different, either owing to the fact that the variability of the associated variates is different, or because the accuracy required is different. If no use is made of the relations between the different variates, such multi-phase sampling is equivalent to taking samples of different sizes for the different items.

First-phase information may also be used as supplementary information in order to improve the accuracy of second-phase information, by the same methods, ratio and regression, that are applicable where supplementary

information on the whole population is available. Thus, in a crop estimation survey based on farms as sampling units, a relatively large sample of farms may be taken for the determination of the acreage of the crop, and the yields may be determined on a sub-sample only of these farms.

If the first-phase information is collected prior to the second-phase information the first-phase information may be used as a basis for the sub-sampling process, *e.g.* by stratification of the first-phase units for the selection of the second-phase sample, with or without the use of a variable sampling fraction at the second phase.

It will be noted that in both these latter applications of two-phase sampling the methods followed are the same as those adopted in ordinary single-phase sampling, the population being replaced by the first-phase sample ; but since the first-phase information is not known for the whole population it is itself subject to sampling error, and this must be taken into account when estimating the sampling errors of the estimates of the second-phase variates.

Multi-phase sampling differs structurally from multi-stage sampling in that in the former the same sampling units are used throughout, whereas in the latter a hierarchy of sampling units is used. Multi-phase sampling may be combined with multi-stage sampling. In a scheme for the estimation of the acreages and yields of agricultural crops, for example, a two-stage sample of farms and parishes may be taken for the estimation of acreages, and a sub-sample of these farms may be taken for the estimation of yields.

3.13 Balanced samples

If the average value of some quantitative character of the units, such as size, is known for the whole population, it is possible, provided the sizes of the individual sampling units are known, to select a sample in such a manner that the average size of the selected units is equal to the average size of all the units of the population. Such a sample will only be satisfactory if it is otherwise equivalent to a random sample, in which case it may be termed a *balanced sample*.

Balance may be employed in conjunction with stratification for some other character. In this case balance may be effected either for the whole population, or for each of the strata separately. The latter course should only be adopted if the number of units selected from each stratum is moderately large : otherwise undue restrictions will be placed on the sample which will result in the selection of a sample which is not otherwise equivalent to a random sample. On the other hand, when the strata are balanced separately more accurate estimates of the separate strata means and totals will be obtained, and the accuracy of the estimates of the overall population means and totals may also be somewhat improved.

Balancing for a known quantitative character provides an alternative to stratification by size-groups in this character. Balancing, however, will only be effective if the differences in the quantity or quantities under investigation

are approximately proportional to differences in the known character, whereas stratification by size-groups will take account of any type of relationship. As will be seen in Chapter 7, the estimation of sampling errors is simpler in the case of a stratified sample.

The increased accuracy resulting from balancing can equally well be obtained, at the expense of some additional computational labour, by adjusting the results of an unbalanced sample by the use of regression in the manner explained in Chapter 6. Since the additional labour of adjustment is nearly proportional to the number of variates under investigation, the advantages of balancing as opposed to regression increase as the number of variates increases.

Balancing can also be carried out for a character which is inherently qualitative, but which for the sampling units actually employed acts as a quantitative variate because the sampling units are themselves aggregates of smaller natural units. Thus if a sample of a human population composed of two different races is being taken, the sampling units being administrative areas containing numbers of individuals, the sample can be balanced for the percentages of individuals in the two races. If the individuals were the sampling units then balance would be equivalent to stratification by races.

A balanced sample is best selected by using a process of replacement. In the first place a random or stratified random sample of the required size is selected, record being kept of the order of selection. The average value of the known quantitative character is then calculated for the sample. This will, in general, not be equal to the average for the population, indicating lack of balance. A further sampling unit is then drawn, and compared with the first unit of the original sample. If balance is improved by substitution of the new unit, this is done, otherwise the original unit is retained. The process is then repeated for the second and following units of the original sample until an adequate degree of balance is attained.

The selection of a sample balanced for more than one character presents more difficult problems, and will not be discussed here.

Balance, in the cruder form known as *purposive selection*, was at one time extensively used in sample censuses and surveys. No rigorous rules of selection were followed, however, with the result that many purposively selected samples were by no means equivalent to balanced random samples. Thus it frequently happened that the selection was confined to sampling units having values of the known quantitative character near the average. Clearly in such samples the variability of the known quantitative character, and of any other characters closely correlated with it, will be considerably less than the true variability in the population. The sample may also be unrepresentative in other ways.

Purposive selection was often used in an attempt to avoid the necessity, which was otherwise apparent, of employing reasonably small sampling units. Thus Gini and Galvani (1929, A) selected a sample from the Italian Census Data of 1921 which consisted of all the returns of 29 out of the 214 *circondari* into which the country is divided, using seven control characters. Agreement of

the average values of other characters in the sample and population was poor, and that of the frequency-distributions of such characters was even worse. The real weakness here is the use of excessively large units, though even with smaller units the use of purposive selection without rigorous rules of selection is always liable to give unsatisfactory results. There is, moreover, no means of judging its reliability.

For these reasons purposive selection has ceased to be extensively used, and in modern sampling work it has largely been replaced by more thorough application of the principles of stratification, etc. Provided proper attention is paid to the process of selection, however, there is no fundamental objection to balanced samples. These have a certain limited usefulness in some types of census and survey work, though it must be recognized that the need for the subdivision of the population into an adequate number of sampling units is in no way obviated by balancing for one or more quantitative characters.

3.14 Systematic samples from areas

A common method of sampling material continuously distributed either in space or time is to take sampling units distributed at equal intervals over the material. The chief application in census and survey work is in the sampling of land areas. When maps are available the sampling units can be located by superimposing a grid of points, frequently of square, or nearly square, pattern. Such a sample may be termed a *systematic area sample*.

A systematic area sample differs from a systematic sample from lists mainly in the spatial distribution of the sampling units over the material. Most lists do not correspond at all exactly, except for major groupings, to any physical distribution, and a systematic sample from a list therefore usually approximates much more closely to a random sample than does a systematic area sample. Different methods of estimating the sampling error are therefore appropriate in the two cases.

In general, provided there are no periodic features, a systematic area sample will be rather more accurate than a stratified random sample (with one unit per stratum) from strata consisting of rectangular blocks (or *cells*) whose centres are situated at the systematic sampling points. In material in which the variation is of a continuous nature it is impossible to make any accurate estimate of the sampling error without taking supplementary sampling points, though if there are no periodic features an upper limit can be obtained.

If the regions near the boundary are likely to differ from the remainder of the area, as may be the case if the boundary is a natural one, such as a sea coast or a mountain range, it will be best, after locating the sampling grid at random, to demarcate the bounding lines of the cells, and sample at random the area which is not covered by complete cells, dividing it into equal or approximately equal areas and locating one sampling point at random within each of these areas. It will be convenient, if possible, to make these cell areas equal in area to those of the sampling grid, since equal weight will then be

given to all sampling points. The same method of dealing with boundaries can be used if the sampling is random within rectangular cells.

Systematic sampling is entirely unsuited to material which has periodic features, but apart from this will generally provide a satisfactory method of area sampling. It has the advantage over stratified random sampling from blocks that the location of the sampling units is simpler and the results obtained provide rather better material for the construction of maps, etc. As in systematic sampling from lists, however, the responsibility for the judgment that the material is such that systematic sampling will give satisfactory results rests with the investigator.

3.15 Line sampling

In the sampling of areas certain types of information can be ascertained almost as easily for all the points on a line as they can for a set of isolated points or areas. In such cases sets of parallel lines or strips may be taken as the sampling units. In stratified random line sampling, the area is divided into rectangular blocks of convenient length and of such a width that two selected sampling lines are included in each block, their location within the block being random. If an exact estimate of the sampling error is not required, only one line need be included in each block, with correspondingly smaller blocks. In systematic line sampling the sample is made up of lines at equal intervals.

Line sampling provides an alternative method to point sampling (Section 3.9) for determining the proportions of a given area which are of different types. These proportions are given by the ratios of the aggregate of the intercepts of the different types. In area determinations of this type, whether by line or point sampling, systematic sampling can usually be adopted. The method is useful both for area determinations on the ground in undeveloped country and for the determination of areas from maps and aerial photographs. The method has been much used for forest surveys, where it is known as "cruising."

If, instead of obtaining information for all points on each of the lines, a sample of such points is taken, the sampling becomes two-stage. If the lines and the points on them are both evenly spaced the sampling is equivalent to systematic point sampling.

Line sampling of a somewhat different type is also used in order to determine the acreage of agricultural crops in areas which are well provided with roads. A route is chosen which covers the whole area as adequately as possible, and the lengths bordered by the different crops are measured. A car fitted with a special milometer can be used for this purpose. Estimates of the yield near harvest time can also be obtained in a similar manner, by stopping the car at every x th mile of a given crop, and cutting and harvesting a small area of the crop, the area being selected in some systematic manner, such as entering the crop a given number of paces at right angles to the road.

Line sampling of this type does not provide an unbiased sample, since

roads are by no means randomly located with regard to agricultural crops. The results from surveys following the same route in successive years may well be comparable, however, and with calibration by more exact methods from time to time, road cruising may provide a satisfactory method of making rapid and inexpensive surveys. Similar methods based on tracks are possible in areas with only a sparse road network.

3.16 The principle of the moving observer

If counts are required of a collection of individuals who are moving about, the ordinary methods of sampling can only be applied with difficulty. Thus, to determine the number of people in a crowded street by ordinary methods would require the demarcation of a number of small areas in the street and the counting of the number of people on each of these areas. The counts need not necessarily be simultaneous, but for any one area the number of people present at a given moment has to be counted. Unless photographic methods are available, or the areas are very small, such counts are extremely difficult, since individuals are continually moving into and out of the areas and are also moving about within them.

Equally it is no use stationing observers at fixed points with instructions to count passers-by. The number of people in a street will depend not only on the numbers passing fixed points but also on the velocity of movement up and down the street. If all exits and entrances to the street are covered, and there are no people in the street at the start of the counts, the number present at any subsequent time can be determined from counts that are continuous and without error. In practice, however, errors in counting usually result in cumulative errors which invalidate the results. Thus it was found impracticable to determine the numbers in a department store by posting people at the doors to make counts.

These difficulties can be overcome by using moving instead of stationary observers. To obtain an estimate of the number of people in a street, the observer traverses the street in one direction, counting all the people he passes, in whichever direction they are moving, and deducting all the people who overtake him. He then re-traverses the street in the opposite direction, moving at the same speed and counting as before. If this is done the average of the two counts gives an estimate of the average number of people in the street during the time of the counts. If people are mostly moving in one direction the count in this direction will be reduced, but the count in the opposite direction will be correspondingly increased. In practice the deductions required for those overtaking the observer can be kept small by moving at a speed greater than that of the majority of the crowd.

This method was used to estimate the numbers of people in streets, shops, etc., at different times of the day, in order that the adequacy of the provisions for public air raid shelters might be tested. It was found that very dense crowds in streets and shops could be estimated with surprising ease. Crowded streets

were dealt with by teams of two or more observers, moving down the street in a transverse line, with each observer counting the people between him and the next observer. In the large stores the floor was divided into areas which were assigned to the different observers.

The method is of general application. It can be used, for example, to assess populations of insects or animals in a state of movement, provided all individuals can be readily seen, and provided the passage of the observer does not itself influence the movement of the individuals.

3.17 Interpenetrating samples

It is often advantageous to take two or more independent samples of a given population, using the same sampling procedure for each sample. Such samples are called *interpenetrating samples*.

Interpenetrating samples are of value if the survey or census has to be carried out by successive stages. This is frequently necessary when preliminary results are required quickly. Thus in the 1942 Census of Woodlands of England and Wales, described in Section 4.25, it was necessary to obtain a preliminary estimate of the timber content with very limited field staff within six months of the initiation of the survey. The work was therefore planned in two interpenetrating samples. Before the field work was commenced, each unit of the first sample was subdivided into two, since it became apparent that the whole of the first sample could not be completed in the allotted time. This further subdivision itself created two interpenetrating samples of a special type. By means of this procedure it was possible to provide a preliminary estimate of the total timber content of the whole country by the time it was required.

An incidental advantage of interpenetrating samples is that separate and independent estimates of the characteristics of the population are furnished. The agreement of such estimates is often more convincing to the layman than any statement of the sampling error.

Interpenetrating samples have a further use in that the different samples can be assigned to different investigators. Comparison of the results provides a check of the investigators against one another.

To perform the functions outlined above, each of the interpenetrating samples must itself provide an adequate sample of the material and must be comparable with the other samples—in other words the samples must be really interpenetrating. If this is not the case the comparisons between the different samples will be subject to relatively large errors. If, for instance, they are used to test differences between different investigators, the information obtained will be of insufficient accuracy to be of any real use. Equally, if one sample is used to provide a preliminary estimate, this estimate may well not attain the required degree of precision. Thus in an agricultural survey stratified by areas such as counties, the division of the counties into two groups, with each of the samples confined to one group only, would not be likely to give

a satisfactory pair of interpenetrating samples. The separate samples would be subject to variation between counties, and would therefore be considerably less accurate than the combination of the two, from which variation between counties is entirely eliminated. The proper use of interpenetrating samples therefore necessitates increased expenditure on travelling.

3.18 Sampling on successive occasions

The types of sampling so far discussed are appropriate to a census or survey carried out on a single occasion, with the object of determining the characteristics of the surveyed population at or about a given point in time. If the population is subject to change, a survey carried out on a single occasion, however accurate, cannot of itself give any information on the nature or rate of such change. In certain types of population extraneous sources of information, such as registrations of births and deaths, may be relied on to provide information on the changes which the population is undergoing. Even in such cases the census must be repeated at intervals, both because of inaccuracies in the extraneous information, which may lead to a gradual accumulation of errors, and also because the information is rarely of such a nature that all aspects of the original census or survey can be kept up-to-date. Registration of births and deaths, for example, coupled with figures for immigration and emigration, will furnish data for the revision of the total of the population but will not enable changes in the population of separate towns and districts to be determined.

In many cases no such extraneous information on the changes that are taking place is available, and in such cases provision must be made for periodical re-survey if up-to-date information is required. A number of alternatives then present themselves :—

- (1) A complete census or survey may be repeated in its original form at intervals.
- (2) A sample census or survey may be repeated at intervals, a new sample being selected on each occasion without regard to previous samples.
- (3) A sample census or survey may be repeated on the same sample.
- (4) Part of the sample may be replaced on each occasion, the remainder being retained. If there are a number of occasions a definite scheme of replacement may be followed, *e.g.* one-third of the sample may be replaced, each selected unit being retained (except for the first two occasions) for three occasions.
- (5) A re-survey of a sub-sample of the original sample may be made. In the case of a complete census this is equivalent to a re-survey of a sample of the whole population.

The following terms are suggested for the last four alternatives :

- (2) independent samples ;
- (3) fixed sample ;
- (4) partial replacement ;
- (5) sub-sample. It will be noted that independent samples are formally

equivalent to interpenetrating samples, a fixed sample is formally equivalent to the observation of different characters (variates) on the same sample, and a sub-sample is formally equivalent to a two-phase sample. Only partial replacement has no formal equivalent in the types of sampling already described. These equivalences are of importance in that the methods of estimation will be the same for formally equivalent sampling processes.

The relative advantages of the various types of procedure depend on the relation between the variability of the units and the variability of changes in these units as well as on the relative importance of information on the population means and on the changes in these means. If, for instance, the units are very variable but the changes of all units are similar, accurate information on change can most easily be obtained by re-survey of a fixed sample of units ; provided always that proper provision is made for new entrants to the population, and for the elimination of the disturbance which results from the extinction of selected units. If, on the other hand, information on the population means is of paramount importance, partial replacement or a sub-sample will usually be preferable. A more detailed discussion, in terms of the errors to which the various estimates are subject, is given in Section 8.8.

There are two further points which must be borne in mind in connection with sampling on successive occasions. Firstly, repeated re-survey of the same units may be inexpedient, since resistance to the provision of the necessary information may be engendered, and secondly, repeated re-survey may result in modification of these units relative to the rest of the population. This can arise in many ways. In a survey of agricultural practice, for instance, visits to farms may result in the farmers concerned improving their practice through advice from the investigators : advice which, if asked for, can scarcely be refused. A more subtle example is provided by the 1942 Census of Woodlands. In this census it was considered that if the subsequent fellings and replantings were recorded on the sample areas then surveyed an adequate measure of the changes in woodland throughout the country might be obtained over some considerable period of time. It has since been suggested that the amount of felling on the sample areas may have been affected by the fact that survey information was available on these areas and not on others, with the result that these areas have been more intensively exploited.

3.19 Composite sampling schemes

Simplicity and uniformity of sampling procedure is obviously in general desirable, but there are occasions on which different methods of sampling are required for different parts of the population. In sampling a human population, for instance, some form of area sampling may be most suitable for the rural parts of the country, whereas some form of stratified random or systematic sampling based on lists of houses may be best in the towns. There is, of course, no objection to the use of such composite schemes, provided each part fulfils the requirements of good sampling procedure already laid down.

3.20 Combination of complete census and sample survey

It sometimes happens that a complete enumeration of the population can easily be made, but that detailed information on the individual units of the population can only be obtained by sampling methods. In such cases a complete enumeration will often be of value as a frame for the sample census. Thus in a census of a human population, a complete enumeration consisting of lists of households and of numbers in each household can be made. A sample of houses can then be visited by investigators so as to obtain details regarding the age, sex, etc., of the occupants. Such a sample census will not only serve to provide the required detailed information, but will also provide a partial check on the accuracy of the complete enumeration. It will not, however, provide a check on omissions from the lists of households. To carry out such a check it will be necessary to take a further sample of properly defined areas, checking that all the households in the sample areas have been included in the full census returns.

A complete census, even if it is very inaccurate, is also of the greatest use in planning a more accurate sample census. In a sample census of a human population, for instance, some knowledge of the relative sizes of different towns and villages, and of the density of population in rural areas, is essential for the proper allocation of resources. Similarly in a census of agriculture, knowledge of the amount of cultivated land in different parts of the country is necessary if excessive survey of largely uncultivated areas is to be avoided.

The information provided by an inaccurate complete census can also be used to improve the accuracy of a subsequent sample census, by the methods applicable to supplementary information which will be given in Chapter 6. Here, however, we must proceed with caution. If, for instance, a complete census of a human population consistently underestimates the population of villages of all sizes by about 10 per cent., the sample census will determine the amount of the underestimation and a common adjustment can be made. If, however, small villages are underestimated by 20 per cent. and large villages by 5 per cent. the application of a common correction will result in the underestimation of the population of small villages and the overestimation of the population of large villages. This distortion will be avoided if separate corrections are calculated for small and large villages. Unfortunately it is not always possible to be certain that all potential disturbances are taken into account. These differential inaccuracies are particularly troublesome in that they tend to be associated with the administrative areas for which separate results are required. The overall results, however, will not be materially affected by differential inaccuracies of this kind if the methods of estimation given in Chapter 6 are followed.

PRACTICAL PROBLEMS ARISING IN THE PLANNING
OF A SURVEY

4.1 Questions requiring consideration

The practical problems encountered in the planning of a sampling investigation vary greatly with the type of material and the nature of the information that is required. We shall here only concern ourselves with problems arising in the conduct of censuses and surveys, such as are required in the study of human populations, economic institutions, and agriculture. The sampling of batches of material, industrial products, etc., which is necessary in manufacturing processes of all kinds, and is broadly categorized by the term *quality control*, presents rather different problems which are not discussed in this book. The sampling problems encountered in biological research are also omitted from the discussion.

The questions that require consideration at the planning stage of censuses and surveys may be broadly classified as follows:—

- (1) Specification of the purposes of the survey.
- (2) Definition of the population, types of institution, or categories of material to be covered by the survey.
- (3) Decision on the nature of the information to be collected.
- (4) Decisions on the method of collecting the data, whether by interviewers, investigators, mail, etc., and methods of dealing with non-response.
- (5) Choice of frame, or construction of a frame if none is available.
- (6) Choice of sampling unit and type of sample, whether stratified, multi-stage, etc., determination of size of sample required, and method of selection.
- (7) Decision on whether the survey is to be an isolated one, undertaken without intention of repetition, or is to be planned with a view to repetition at intervals.

These questions cannot be considered in isolation one from another. To a greater or less extent any decision taken on one question will influence the decisions that should be taken on the others. They should therefore be resolved jointly, or if independent decisions are made these should at least be regarded as tentative and subject to modification until the plan as a whole has been finalized.

Nor can the correct decisions be arrived at without considerable knowledge of the nature of the material to be covered, particularly its variability, both as a whole and within and between strata of various types. Knowledge is also required of the ways in which it is practicable to collect the required information with the necessary accuracy. If prior knowledge in these matters is not available a *pilot or exploratory survey* will be necessary. Even if there is

adequate knowledge of the statistical properties of the material, pilot surveys are frequently advisable in large-scale surveys in order to test and improve field procedure and schedules, and to train field workers.

Questions arising under heads (1), (2), (3), (4) and (7) of the above list are common both to complete censuses and surveys and to sample censuses and surveys. Even here, however, the problems encountered differ considerably in the two cases, owing to the greater scope for the collection of detailed information and the execution of complicated observations by the sampling method.

The determination of the items on which information is to be collected, the degree of detail to be attempted, and the ways in which the information can best be obtained, often constitute the most difficult and crucial part of the planning of a survey. No amount of care in the planning of the sampling or skill in the analysis will compensate for failure in this respect. A survey in which the information collected does not adequately cover the field to be investigated at the best provides a partial and incomplete picture, and at the worst may be irrelevant or actively misleading.

Careful consideration must therefore be given at the outset to the purposes for which the survey is to be undertaken, the type of information it is proposed to collect, and the uses to which the information obtained will be put. In the case of large-scale surveys, which are likely to provide information that will be of value to a number of different organizations or government departments, a detailed statement on these points should be prepared. In this way those who are likely to want to make use of the results of the survey will be fully apprised of its nature, and can if necessary make suggestions for modifications before the survey is begun.

The statistician who will ultimately be responsible for the analysis and presentation of the results should, if possible, be selected and appointed at the planning stage. Similarly if the advice of a statistical expert is to be sought, this should be done, in the first instance, at the planning stage. This rule applies even in the simplest types of census. It frequently happens that such censuses are undertaken without any prior consultation with a statistical expert, whose advice is only sought when the results have been collected and the stage of analysis is reached.

4.2 Definition of the population

The categories, or types of material, which require to be included in a survey, and its geographical scope, are conditioned in broad outline by the purposes of the survey, and by administrative and research requirements related to these purposes. Within this broad outline, however, there is often a certain amount of latitude, and careful consideration should therefore be given to the inclusion or omission of marginal categories, particularly those on which the collection of information is likely to be specially difficult, or for which an adequate frame is lacking. By excluding unimportant marginal categories

the task of collecting the information may often be very materially simplified, without seriously reducing the value of the results.

A census of the human population residing in a given territory, for example, should ideally include all individuals present in that territory at a particular moment, and in simple censuses an attempt is usually made to attain this end. It is often, however, difficult to obtain information on certain minor categories, such as nomads. These difficulties occur even in a complete census, but are often more marked in the case of a sample census. The question of whether such categories may be omitted entirely without serious loss should therefore be considered.

The matter becomes of even greater importance when a human population census requiring the collection of detailed and complicated information is undertaken, using skilled investigators making visits to individual members of the population. In such cases visits to members of the population with a permanent residence, even if they are absent from their residence at certain times, are relatively simple, but it is far more difficult to cover the floating elements of the population. The conduct of such a census becomes very much simpler, therefore, if these latter elements can be omitted.

In a similar manner, in the case of an agricultural census, the determination of the areas of the various crops might ideally require that all areas of the crops grown within the boundaries of the territory should be included. It may, however, be possible to exclude small areas, such as those found in gardens and holdings of very small size, without seriously reducing the value of the information. The agricultural censuses of England and Wales, for example, which are based on returns from farmers, exclude all agricultural holdings of less than one acre, and do not attempt to take account of crops grown in private gardens or allotments.

The question of whether or not minor categories should be included depends mainly on the purposes for which the information is required. A case is sometimes made for the inclusion of certain categories on which the information is intrinsically of little interest in order to ensure comparability with the results of previous censuses or surveys, or with the results of parallel surveys in other countries. Comparability within and between statistical series is obviously desirable, and lack of it can seriously reduce the value of the results, and also increase the labour of statistical analyses and the danger that those unfamiliar with the details of the various sources of information may draw wrong conclusions. Nevertheless when introducing a radically new method of collecting information, such as replacement of a complete census by the sampling method, excessive weight should not be given to past practice. It should not be forgotten that so-called complete censuses are often in themselves subject to errors of various kinds, including lack of completeness, and that such errors are often a greater source of disturbance to comparability than the omission or inclusion of a few minor categories. If there is any serious doubt whether a given category should or should not be included this may be regarded as *prima facie* evidence that the category in question differs in

essentials from the other more important categories. Consequently, if it is decided that the category should be included, the results should be kept separate so that they can be summarized separately and eliminated from the final estimates if required, or given special treatment in these estimates. If this is done for the first one or two surveys of the new type, comparability with previous results will be ensured, without preventing the omission of the category in subsequent surveys if this ultimately appears desirable.

The arguments in favour of the adoption of identical definitions in different countries in which conditions are radically different are even less strong. Categories which are of very minor importance in one country may be of great importance in another. Decisions as to their inclusion or omission should be taken primarily on the grounds of their importance in the country which is being surveyed, without undue regard to definitions designed to ensure formal uniformity of world statistics.

In many cases in which complete omission of unimportant categories would not be justified, they can be very conveniently dealt with by some special sampling procedure, which may be multi-phase, or may be of an entirely different type with different frame and sampling units. Thus in a human census, certain of the simpler items of information, which can be reliably furnished by neighbours or other members of the household, may be collected for absentees abroad, or a sub-sample of these absentees may be taken for a follow-up enquiry by more intensive methods. Nomads may be dealt with by instituting a supplementary sample census to deal only with this category of the population.

In a sample survey the frame adopted contains its own implicit definitions of the categories of material to be covered. If a category is not included in the frame it will either have to be omitted entirely or special steps will have to be taken to supplement the frame. Definitions of the population should therefore be considered in conjunction with the choice of frame.

4.3 Determination of the details of the information to be collected

The detailed problems which arise in deciding what information is necessary and how it can best be obtained vary widely in surveys covering different fields of enquiry and according to whether the results are required primarily for administrative or for research purposes. Full discussion of any particular case necessarily requires extensive knowledge of the subject as a whole and of the particular questions at issue, and would be out of place here, but there are certain general points which may be mentioned.

The basic problem is essentially that of the selection of the most relevant items of information or types of observation from all those which it is practicable to collect and which might conceivably have a bearing on the matters under investigation. This selection must be such that a coherent whole is obtained which covers the required field adequately, or if this is not possible at least provides information on some relevant part of it.

This basic problem is essentially the same in complete censuses and sample censuses and surveys, but the problem is more complex in the case of sample censuses and surveys, since the items of information that can be collected and the observations that can be made are themselves more complex and varied.

The best way of arriving at a satisfactory solution of this basic problem is usually as follows. In the first instance, the details of the information required to deal with the problems originally propounded are determined. The question is then considered whether there are any related problems of importance on which this information, possibly supplemented to some extent, would throw light. If this is the case the supplementary items of information required for the full elucidation of these additional problems should be determined. With the whole field mapped out in this way, the practicability of obtaining the necessary items of information covering any given set of problems can be considered, and final decisions taken in the light of the relative importance of these problems and the total load which it is considered expedient to place on the investigators and respondents in a single survey.

The details of this process vary greatly in different types of survey, but the general principle to follow in all types of survey is to see that the items of information collected form a rounded whole covering a definite subject or coherent group of subjects.

This principle is of particular importance in surveys of the questionnaire type on human populations, whether the questionnaires are filled in by the respondents themselves, or the information is elicited by field investigators. Accurate information can only be obtained in such surveys if full and willing co-operation of those providing the information is obtained. The survey must therefore have a clear purpose which can be explained to the respondents, and the questions asked must be relevant to this purpose. If additional questions dealing with unrelated subjects are included, or if the questions relating to the main enquiry seem trivial, and do not cover aspects which appear of importance to those providing the information, the survey will cease to appear as a serious enquiry into a particular subject, and will meet with unfavourable reactions, summed up in such terms as "snooping."

The matter is of importance even in enquiries which require the collection of factual information by observation and measurement by the investigators themselves, without any co-operation from respondents. If the field investigators are not imbued with a sense of the importance of their enquiry, and are overloaded with the collection of miscellaneous data, they will not give of their best. Occasionally information may be sought on points unconnected with the main survey if it is urgently needed, and considerable expense is thereby saved, *e.g.* in travelling, but this should be avoided as far as possible.

Occasionally, in cases in which a questionnaire would otherwise be unduly long it may be possible to split it into parts, obtaining information on one group of items from one set of respondents, and on another group of items

from a second set. Certain basic items of information will be required from all respondents, and the two sets will form a pair of interpenetrating samples. The sampling has also a two-phase structure, the basic information acting as first-phase information for both sets. If this procedure is followed, however, the relationships between items of information in the two groups can only be studied for strata or other suitable domains of study, and not for the individual respondents.

Certain items of information are often required in order to ensure the proper interpretation of other items. Thus, for example, if housewives are being asked whether they prefer coal, gas or electric cooking, and the reasons for their preferences, it is essential to ascertain in some detail what experience they have had of methods of cooking other than the one they are now using, including the type of apparatus used. If this is not done the answers may be more an indication of the effectiveness of an advertising campaign in favour of one of the methods, or a condemnation of antiquated pieces of apparatus, rather than any reflection of the true relative merits of the different methods.

Information is also often required on items which, though not of primary interest, will act as supplementary information and thereby enable the precision of the results to be increased by the appropriate methods of estimation.

In reaching the decisions on the type of information required, both in broad outline and in detail, it is absolutely essential to work in collaboration with experts on the subjects which it is proposed to cover. If research or administrative experience in the subjects to be covered is lacking, it is fatally easy when designing a survey to omit some vital items of information. A simple instance of such omission is provided by the 1921 and 1931 Population Censuses of the United Kingdom. In these censuses information on age of mother at marriage, and total number of children born, which had been obtained in the 1911 Census, was not asked, with the result that the value of the information provided by these censuses for studies on changes of fertility of the population has been very seriously reduced. As a result of this lack of information it was felt to be necessary to institute a special Family Census in 1946 (Section 4.10). In this instance it can scarcely be that the need for this information was wholly overlooked, but insufficient weight must clearly have been given to this aspect of census information.

In addition to direct collaboration with experts in the various subjects, the plans for the survey should be circulated at all stages of development to the various organizations and individuals who are likely to be interested in the results. This will usually result in requests for the collection of supplementary items of information, some of which may not be necessary for the purpose for which the survey was originally planned but which will enable the results to be used for other purposes. In this way the usefulness of the survey may often be considerably increased. On the other hand, the danger of overloading the survey with the collection of miscellaneous items of information must be guarded against, and all requests should therefore be very carefully reviewed.

4.4 Inter-relations of groups of natural units

If the physical inter-relations between the members of groups of natural units of the material under survey are of interest, or if information is required for groups of natural units as a whole, then information must be collected for such groups as a whole, or at least for pairs of units from such groups. Thus if the inter-relations between the different members of a household require to be studied, it is essential to have information for pairs of individuals belonging to the same household, and it is usually best that the information should cover the whole of a household. This can be ensured by using households or dwellings as sampling units.

Another type of natural aggregate for which it is often important to obtain results as a whole is that provided by towns, villages, etc., and, in agricultural surveys, homogeneous geographical areas. This often calls for the adoption of multi-stage sampling, the natural aggregates forming the sampling units at the first stage, even in cases where the use of single-stage sampling is otherwise preferable. Thus in a survey of a human population it may be of considerable interest to contrast the results for individual towns of differing types, and to study the inter-relations existing within a single town, even when there is no need for all the towns of the country to be covered.

Similarly, if inter-relations between the behaviour of the same individuals or other natural units at different times are of interest, the survey must be designed so as to provide information covering an adequate period of time. Thus in an investigation into hours of sleep of children, it is of little value to determine the amount of sleep of a sample of children each for a single day. Such data will throw no light on the question of whether children who have a short period of sleep on a particular day tend also to go short of sleep on other days or are able to make up for this short period by longer periods on preceding or following days. In the same way, studies of nutrition in which the intake of food is determined for each individual for a single day only, although they will show whether a group as a whole is under-nourished, are incapable of revealing the degree of variation in under-nourishment between individual and individual, since individuals going short of food on a particular day may make up for such deficiencies, in whole or in part, on succeeding days.

We have stressed this point at some length because there has been a tendency, in surveys on human populations of the questionnaire type, to take the relatively easy course of asking those interviewed about occurrences which are still fresh in their mind, *e.g.* what happened on the previous day. This course is followed for various reasons. It may be considered that information provided about earlier occurrences will be inaccurate, or that there is a danger of overburdening respondents if an attempt is made to cover too long a period; or the object may be to save interviewers the trouble of repeated visits which might be required to cover a period of time accurately. Actually it has been found that the use of a very short period does not necessarily lead to accurate average results: in certain circumstances there may be a tendency on the part of

respondents, either consciously or unconsciously, to telescope events, and report them as happening in the given period when in fact they happened earlier. Thus a survey of crockery breakages made by asking what breakages occurred over the past week led to an entirely excessive estimate of the amount of breakage, whereas a similar survey asking for breakages over the past year gave results which checked well with production figures and the domestic stocks (Box and Thomas, 1944, D, discussion).

4.5 Practicability of obtaining the required information

So far we have been considering the problem of determining what items of information are required in order that the purposes of the survey may be fulfilled. Each item must, however, be considered in the light of the practicability of obtaining it. If the information is to be furnished in response to questions, the points for consideration are whether the respondents are sufficiently informed to be capable of giving accurate answers ; whether, if the provision of accurate answers involves them in a good deal of work, such as consulting previous records, they will be prepared to undertake this work ; whether they have motives for concealing the truth, and if so whether they will merely refuse to answer, or will give incorrect replies. If the information is to be obtained by observation or physical measurement, the points for consideration are whether the observations are such that they are within the competence of the investigators or other individuals who will be required to undertake them ; whether they will make excessive demands on the time of the investigators or others, or require excessively expensive apparatus ; and whether the owners of the surveyed material will permit the observations to be made.

Considerations of this kind will inevitably lead to modifications of what would otherwise be considered an ideal scheme. Nor can general answers be given, even within the limits of a particular field of enquiry. In countries such as the United Kingdom, for example, there is no reason to suppose that any large amount of inaccuracy is introduced into the returns of the population censuses by deliberate mis-statements. In countries not accustomed to population censuses fear that the information will be used for such purposes as taxation or conscription may lead to considerable inaccuracies. Similarly in crop-sampling work the use of small sample areas may be quite satisfactory with certain classes of field worker, but, as is shown by Table 2.5, is entirely unsatisfactory in other cases.

When the ideal requirements cannot be fully met it is sometimes possible to include other items of information, observations, or physical measurements, which, owing to their high correlation with the quantities which it is desired to determine, will serve as more or less adequate substitutes for these quantities. These substitute measures may be used for purposes of stratification or classification of the data in the final analysis, as for example when the rateable value of a dwelling is used as a substitute for the income of the household

occupying it ; or they may be substitutes for measures of quantities which themselves require assessment, as for example the use of eye estimates in place of direct measurements of the yields of a standing crop. The efficiency of such substitute measures can only be properly judged by a proper statistical investigation of the relations between them and the quantities for which they are substitutes. In the case of substitutes for measures of quantities that are to be assessed, some method of calibration is essential if objective estimates of the original quantities are to be obtained. (The calibration of eye estimates is discussed in Sections 6.15 and 7.14.)

It will inevitably happen in certain cases that information which is of considerable importance will prove to be unobtainable, or unobtainable with sufficient accuracy. When such a situation arises it must be squarely faced. There is at times a tendency to attempt to collect information which, because of its nature, cannot be obtained with the necessary accuracy, and then to condemn the survey method in general because the results are of little value.

This, however, does not mean that the collection of difficult items of information should not be attempted. The sample survey procedure, because it makes possible the use of skilled investigators working on a relatively small sample, is frequently capable of eliciting reliable information on points which it would be quite impossible to include in a general enquiry. The fact that the enquiry *is* on a small sample, if known to the respondents, frequently makes them willing to give information which they would certainly not be prepared to give if the enquiry were general. In such cases it is important that the investigators should themselves be recognized as impartial and disinterested ; in particular they should not be officials of an organization which itself might make use of the information obtained to the detriment of the respondents.

Nevertheless there are subjects on which it is impossible to collect accurate information from a random sample of the population. In certain of these cases information can be collected from a selected group of individuals, *e.g.* individuals with whom social welfare workers are in contact. Information of this type is not necessarily valueless, but it must be clearly recognized that it is not the equivalent of information obtained from a random sample of the whole population, and any attempted generalization of the results will be of limited validity.

Attempts are sometimes made to obtain a sample from such a group of individuals which conforms more closely in certain respects to the population, *e.g.* in classification by age or social class, than does the group as a whole. While this may improve the sample somewhat, it still does not provide the equivalent of a random sample. On the other hand, if the whole of the group is not required, it is usually advisable to apply some rigorous form of selection rather than to permit the workers themselves to select individuals for investigation, as the latter procedure will merely introduce further unnecessary elements of bias.

In cases in which some of the items of information are difficult to collect, multi-phase sampling may be of value. It may, for instance, enable specially skilled investigators to be used for the more difficult items. Thus in a health survey medically qualified investigators may be used on a small sub-sample of a much larger sample on which more general items of information relating to health have been collected. Equally it may be used to reduce the work required to manageable proportions. Thus, in the Survey of Fertilizer Practice soil samples for chemical analysis were taken from one old-arable field, one new-arable field, and one field of permanent grass on each farm, these fields being a sub-sample of all the fields on which information on the use of fertilizers was obtained (Section 4.23).

4.6 Methods of collecting the information

The methods of collecting the information are to a large extent conditioned by the material under survey and the type of information required. Where the alternative possibilities exist, it may be stated as a general rule that observations are preferable to questions, and questions on facts and on past actions are preferable to questions on generalities and on hypothetical future conduct. Thus it is better to inspect a house to see if it shows signs of damp, than to ask the occupant if the house is damp; and it is better to find out what considerations, from among the various alternatives (if any) that presented themselves, governed the selection of the house in which the occupant is living, rather than to ask what type of dwelling—house, flat, bungalow, etc.—is “preferred.”

On the other hand, it is scarcely possible to state any general rule with regard to physical measurements and qualitative observations made by the investigator. Physical measurements are more objective, but qualitative observations are often more capable of summing up the salient features of a complex situation. Thus a qualitative grading by the investigator of the degree of dampness of a house is likely to be more effective than any physical measurements designed to determine the degree of dampness. Moreover, by proper standardization and calibration among investigators qualitative observations can themselves be made objective.

When the information is collected by means of a census form or questionnaire the questions which are to be asked should be considered at the planning stage, since the information obtained will depend on the exact form of these questions. Equally the exact form of any observations and physical measurements which are required should be determined.

Census forms and questionnaires may be designed either for completion by the respondent with little or no assistance from investigators, or for completion by the investigator by the aid of questions put to the respondents. In questionnaires of the latter type the investigators may be instructed to ask questions with a given form of wording, or they may be instructed to elicit information which will provide an answer to the questions of the questionnaire

by enquiry and discussion without adherence to any exact form of words. Both means of eliciting information may be required in the same survey for different items of information.

Census forms and questionnaires designed for completion by the respondent may be delivered and returned by post, delivered by post and collected by an enumerator or investigator, or vice versa, or delivered and collected by an investigator. Use of the post is clearly most economical, and is the method generally followed in censuses and surveys of industrial and commercial organizations, such as censuses of production. In such cases the use of investigators will not normally have any great advantage over the post, either in ensuring more complete response or obtaining more accurate information, though occasionally in local surveys investigators may be used to explain the purposes of the survey and persuade the respondents to co-operate. In population censuses, however, investigators are normally used both in order to ensure the maximum response, and to give assistance where necessary in filling up the forms. Censuses and surveys of small-scale industrial and commercial organizations, and of farms, occupy an intermediate position, and the method used will depend to a large extent on local circumstances.

Attention must be paid to the detailed wording of all questions, even if these are only intended as guides to the investigator. If the question itself creates a wrong impression in the mind of the investigator this will undoubtedly lead to errors, even if additional explanatory notes indicate that something else is really required.

Careful thought must also be given to the order of the questions. If questions are arranged in an orderly sequence the investigator's task is much easier, and the respondent's reaction is likely to be more favourable. This applies to all forms of questionnaire, but is most important in the verbal questionnaire.

In many types of survey it is profitable to give the investigator or respondent an opportunity of making general remarks on special points. This can be done very simply by including a space for observations. Some guidance should be given on the type of observations required. Although such observations do not easily lend themselves to exact analysis they are frequently of considerable value in drawing attention to relevant facts not covered by the questionnaire itself.

The type of investigator to be employed must also be considered. Investigators should have a background knowledge of the subject under investigation, particularly in investigations of the research type. In a technical investigation into housing conditions, for instance, the investigators should have some knowledge of housing construction and of standards normally adopted in good practice. This requirement of technical knowledge in the investigators limits the scope of unspecialized teams of investigators. Such teams are suitable for carrying out *ad hoc* and routine investigations which require only relatively simple questionnaires, but they are no substitute for the more specialized teams required for investigations of a research nature involving technical questionnaires and observations or physical measurements by the investigators themselves.

In surveys requiring any high degree of technical knowledge it is usually best either to use members of existing organizations, or to appoint a small specialized team of technically qualified research investigators. The various surveys into the technical and economic aspects of agricultural practice in England and Wales, for example, are carried out by the staffs of the National Agricultural Advisory Service and the Provincial Advisory Economists. By this means teams of investigators are obtained who are technically qualified and capable of discussing the problems involved with the farmers; at the same time the investigators themselves gain a wider knowledge of the farms of their district which is of value to them in their other work.

4.7 Methods of dealing with non-response

Unless non-response is confined to a small proportion of the whole sample the results cannot claim any general validity. Every effort must therefore be made to reduce non-response to negligible proportions.

Non-response is usually most serious in postal questionnaires. Delays in response can also sometimes be very troublesome, particularly when the results are required quickly. A rigorous system of dealing with failure to respond and delay in response must therefore be instituted at the outset. The first step is to send a follow-up letter, but if this does not produce the required effect, the possibility of using more intensive methods such as telephone calls and personal visits must be considered. These will require a special regional organization.

In censuses of industrial and commercial undertakings in which data on production, sales, labour force, etc. are required for the purposes of economic planning it is usually possible to make the returns compulsory. This is often a help in dealing with a small minority of recalcitrant institutions, particularly if pressure can be brought to bear in other ways, but it is no substitute for full and willing co-operation by the majority. Complete population censuses are usually also made compulsory, and there appears to be no logical reason why sample censuses of the same type should not also be compulsory. While this is little help in dealing with obstinate refusals, since the census authorities are not likely to wish to bring the offenders before the courts, it is an indication that the government regard the census as of importance, and to this extent is likely to act as a persuasive force with the waverers.

In censuses which are to be repeated at intervals it is particularly important to deal vigorously with non-response and delay in response at the outset, as otherwise they tend to increase progressively. If any large volume of non-response persists, or if there is any serious delay in making the returns, it is an indication that something is wrong with the census, which should either be reorganized or abandoned.

In sociological surveys using the interview method the amount of deliberate non-response is usually small. If it is not, the questionnaire and the type of investigator used should be reviewed. Revisits by special investigators

can be tried, but are not likely to be very effective. In technical surveys of agriculture involving interviews with the farmers the amount of deliberate non-response is also usually small, unless the amount of information required is such that it puts too heavy a burden on the respondents.

In sociological surveys, however, initial non-response due to failure to contact the respondent can be very troublesome. There is no proper way of dealing with this except by persistent call-backs. The number of call-backs can often be reduced by enquiring of neighbours when the respondent is likely to be at home, or where he can be found so that an interview can be arranged. Call-backs are also required because the respondent, though willing to give the information, is otherwise engaged at the time of the first call.

The amount of work involved in follow-ups and call-backs can be reduced, if this appears desirable, by taking a sub-sample of those not contacted at the first (or subsequent) call, and weighting up the sub-sample in the final results. In repeated censuses, however, complete follow-ups are advantageous in encouraging better response to later censuses.

4.8 The frame

The whole structure of a sampling survey is to a considerable extent determined by the frame. The methods of survey which are suitable for a given type of material may be radically different in different territories because different types of frame have to be used. Consequently, until particulars of the nature and accuracy of the available frames have been obtained, no detailed planning of the survey can be undertaken. If no frame exists, the construction of a frame suitable for the purposes of the survey may well constitute a major part of the work of the survey.

Frames are subject to various types of defect, which may be broadly classified as follows. A frame may be :

- (1) Inaccurate.
- (2) Incomplete.
- (3) Subject to duplication.
- (4) Inadequate.
- (5) Out of date.

A frame may be termed *inaccurate* if information about the units listed in it or defined by it is inaccurate. The term may also be used to cover the listing of units which do not in fact exist. Thus a ration-card list in which certain women were incorrectly described as married when they were in fact single, or in which certain individuals were included who had died, would be inaccurate in these respects.

A frame may be said to be *incomplete* when certain units of the material are omitted entirely, and be *subject to duplication* when certain units of the material are included more than once. Thus a ration-card list in which certain individuals

were not included, and others were included twice, would be both incomplete and subject to duplication.

A frame may be termed *inadequate* when it does not cover all the categories of the material which it is desired to include in the survey. Thus a ration-card list which did not include the temporary residents in a district would be inadequate for a survey of the population of that district in which it was necessary to include such temporary residents.

A frame, though accurate, complete, and free from duplication at the time it was constructed, may no longer be so at the time it is required for use. Such frames may be said to be *out of date*. Errors of all the first three of the above types may be introduced through the frame being out of date.

These different types of defect have very different consequences in the defects they introduce into the sampling process. Inaccuracy in the frame, in so far as it relates to the selected sampling units, will automatically be discovered and corrected as the survey progresses, and consequently will not invalidate the results. If the information contained in the frame has been used as a basis of stratification, etc. or as supplementary information, inaccuracy in this information will result in somewhat lower accuracy in the results, but the actual accuracy attained will be assessable from the results themselves.

Incompleteness in the frame will not be discovered in the course of the survey itself, and to the extent to which a frame is incomplete the population or material will fail to be covered. Incompleteness is likely to be more serious than it appears to be at first sight, since it is often confined to units possessing some special characteristics, which may in consequence be seriously under-represented in the sample. Duplication has a similar effect, since the duplicated units will have a double chance of being included in the sample. There is the difference, however, that incompleteness cannot be determined or set right by an examination of the frame itself, whereas duplication may under certain circumstances be detected and corrected by such examination, though this will almost always be a tedious operation. If the sampling fraction is large and the degree of duplication is also large, the duplication may come to light in the course of the survey. Thus, with 5 per cent. duplication and a sampling fraction of 1 in 10, two out of every 210 units in the sample will on the average constitute a duplicate pair. With a sampling fraction of 1 in 100, however, only two out of every 2100 units in the sample will constitute a duplicate pair.

A frame which is inaccurate for certain purposes may be incomplete for others. Thus a ration-card list in which some of the single women were described as married would be complete, though inaccurate, if used as a frame for a survey of all women, but would be incomplete if used as a frame for the survey of single women only. Such incompleteness could be remedied by taking a sample covering all women, and rejecting those members of the sample who were found on investigation to be married.

Inadequacy of the frame will usually be known before the survey is undertaken from the specification of the frame itself. Inadequacy can in general

only be dealt with by the construction of a subsidiary frame for the omitted categories.

In actual practice, frames are likely to suffer to a greater or less extent from all of the above defects. It is therefore essential at the outset of the survey to carry out a careful investigation of any frame it is proposed to use, since many defects are not at all apparent until a detailed investigation has been made. Such an investigation will naturally commence with a study of the administrative machinery by which the frame has been constructed and by which it is kept up-to-date, but may also have to include a certain amount of field work.

4.9 Frames suitable for censuses and surveys of human populations

Human populations have a tendency to aggregate in towns and villages, often with very high local densities, which makes any form of area sampling based on maps and plans subject to high variability, unless a very elaborate sampling procedure is adopted. This is most serious if the total numbers are not known, and require to be estimated from the sample, but even the proportions of the population falling in different categories will be subject to substantial errors, since different classes of the population tend to be concentrated in different areas.

Three very different types of survey of human populations may be distinguished. These are :

- (1) Surveys of the census type, requiring the collection of relatively simple facts, but covering the whole population, and capable of giving separate results for small administrative areas.
- (2) Surveys covering the whole population of a country, and capable of giving reasonably accurate estimates for the whole population, and possibly for certain broad subdivisions, but not for small administrative areas. Such surveys often involve the collection of more detailed and elaborate information than do those of type (1).
- (3) Local surveys covering a particular town or rural area, or a few contrasted towns or rural areas, in which no attempt is made to obtain a sample which is fully representative of the country as a whole. Such surveys almost always involve the collection of detailed information by field investigators. They are usually investigations of a research nature, and may be precursors of simplified surveys on the same problems covering the whole country.

Surveys of the first type present relatively simple sampling problems, and relatively complicated administrative problems. The sampling, since it has to cover small subdivisions of the population, must generally be single-stage, usually with stratification and a uniform sampling fraction. Surveys of the third type are also relatively simple ; since only limited areas have to be covered, a one- or two-stage sampling process usually suffices.

Surveys of the second type, however, present much more difficult sampling problems, and also give much greater scope for increase in efficiency by the use of the more elaborate sampling methods. Since results are not required for small areas, administrative or other areas can form the first stage of a multi-stage process, thus enabling the sampling to be concentrated in relatively few areas instead of being spread over the whole country. This condition is absolutely necessary when field investigators are used.

In fully developed areas a good deal of prior information on administrative areas is usually available. This often enables the accuracy of sampling at the first stage, which in general is the stage which contributes most to sampling error, to be substantially improved by the judicious use of stratification, supplementary information, etc. The sampling problems of this second type of survey are discussed in more detail in Section 4.18.

Frames suitable for the sampling of human populations may be broadly classified as follows:

- (a) Lists of individuals in the population, or parts of it, provided for administrative purposes.
- (b) Aggregates of census returns resulting from a complete census.
- (c) Lists of households or dwellings in given areas.
- (d) Town plans.
- (e) Maps of the rural areas.
- (f) Lists of towns, villages, and administrative areas, often with supplementary information of various types.

In the following sections we will give a brief description of the outstanding features, from a sampling point of view, of these various types of frame.

4.10 Frames from lists of individuals

Lists and card indexes of individuals are provided by various administrative activities, such as registration of the population, rationing, or even a recent census. Such lists are only likely to be complete, accurate, and up-to-date if the administrative machinery is very efficient, and there is some definite administrative need for the lists to be kept under constant revision. Most lists of this type cover the whole of a country on a more or less uniform basis, but they are necessarily maintained by local offices, and their accuracy may consequently vary in different parts of the country. Even when a list is sufficiently accurate to fulfil adequately the administrative purposes for which it was designed, it will often be found to have unsuspected defects which make it unsuitable as a sampling frame.

Lists of individuals are not suitable as a frame for sampling households, unless the individuals are grouped by households. If addresses are selected from a list not so grouped, the probability of selection will be proportional to

the size of the household, which is rarely what is required. If such a method of selection is for any reason used, the results must be weighted in inverse ratio to the number of people in the sampled households included in the list (Section 6.16).

Examples of administrative lists are provided by the National Registration and ration-book lists maintained in the United Kingdom. The National Register, which was instituted in 1943, has probably always constituted a reasonably accurate register of the whole population, but in its early stages, at any rate, it was very defective as a local register, owing to the failure of individuals to register changes of address. This defect was later rectified by establishment of joint offices with the food offices, and insistence that any applicant for a new ration book should first have his identity card amended. This, however, did not ensure immediate registration of local changes of address, since new counterfoils were only required if the removal necessitated change of shops. Consequently local changes were often only registered at the time of the regular yearly issue of ration books.

The card index of the ration-book issues necessarily suffered from similar defects. Consequently neither of these registers formed a suitable frame for the sampling of small administrative areas, such as a single food-office district, particularly during the war when movements of population were frequent and considerable owing to air raids. On the other hand, they were and are capable of serving as a reasonably adequate frame for a sample census of the whole population.

The food-office card index was used as the frame for the 1946 Family Census of the United Kingdom. This census was carried out by the Royal Commission on Population, with the object of providing, for married women, information on age, age at marriage, number and dates of birth of all children, and husband's occupation, information which had never previously been collected in full. A sample of 1 in 10 of all the married women (including those widowed or divorced) was taken, by examining every tenth card, and recording the name and address if the card was for a female adult with the prefix Mrs. or with no prefix. Unmarried women selected by this process were requested to mark the questionnaire "unmarried." Questionnaires were dispatched by post, and collected by subsequent visit.

Since there is necessarily a time lag in cancellation of the old food-office card on removal—this is effected by notification from the food office issuing the new counterfoils—special steps had to be taken to deal with removals. This was effected by fixing a "zero" date at an interval prior to the date when the sample was taken. The interval was chosen so as to be somewhat longer than the time taken for notification of change of address to be received at the old office. Thus virtually all duplicate cards corresponding to changes of address prior to the zero date would have been removed. Registrations effected after the zero date were excluded from the sample, and all cancellations bearing a date of re-registration subsequent to the zero date were sampled, the new address being recorded. It will be seen that by this procedure all individuals

entered into the sampling frame once and once only, and the only individuals for whom incorrect addresses were recorded were those who had for some reason delayed their re-registration, or for whom notification of change of address had not yet been received by the old office. This procedure avoided all duplication except in the rare event of excessive delay in notification between offices, while obviating the necessity of any attempt to construct a fully up-to-date non-duplicated index.

4.11 Frames from complete population censuses

A complete census, in so far as it really is complete, will automatically provide an aggregate of forms which includes all the individuals in the area covered by the census. Nevertheless complete censuses, although they would appear at first sight to provide very satisfactory frames, have a number of defects. A complete census by its very nature can only be carried out at infrequent intervals, *e.g.* every ten years, and consequently the frame provided by such a census is for the greater part of its existence badly out of date. The way in which the census information is customarily collected and analysed also tends to reduce its utility as a frame, since the information is not readily accessible, at least in the early stages during which it is being transferred to punched cards. One of the great advantages of the food office register used in the Family Census was that the cards could be consulted and sampled in the local offices without serious disturbance of the office routine.

Many of these disadvantages can be overcome if at the time a complete census is undertaken arrangements are made to construct a proper *master sample* from which further samples can be drawn as required. The sampling unit for such a master sample should be the dwelling, and not the individual or household occupying that dwelling at the time of the census. If dwellings are adopted as sampling units the master sample will have a much greater degree of permanence than would be the case if individuals were used as sampling units. Furthermore, for most purposes a sample of households, and not of individuals scattered over all households, is required.

A complete census will provide a very suitable frame for a simultaneous sampling census in which more detailed information is collected on a sample of the population. This procedure was used in the 1940 census of population in the U.S.A. (Stephan *et al.*, 1940, C). In this census supplementary questions were asked of 1 in 20 of the individuals included in the complete census, at the same time as the main census information was collected. The procedure was thus analogous to two-phase sampling, with the exception that the first phase constituted the whole population.

Since the selection of 1 in 20 individuals for the collection of the supplementary information was done on the spot by the field investigators, certain very rigorous rules had to be instituted in order to avoid bias. The actual procedure adopted was as follows. The census forms each contained lines for 80 individuals, 40 lines on each side. Two of the lines on each side were specially marked.

Five different types of form were used, with the marked lines distributed in the manner shown in Table 4.11.

TABLE 4.11—SAMPLING LINE NUMBERS IN THE 1940 U.S. POPULATION CENSUS,
AND THEIR PROPORTIONS

Style	Proportion	Line numbers			
V	16	14	29	55	68
W	1	1	5	41	75
X	1	2	6	42	77
Y	1	3	39	44	79
Z	1	4	40	46	80

The investigators were instructed to enter the names of each family in a defined order, and to complete all lines of the form before commencing a new form. Actually these instructions were not always adhered to, $3\frac{1}{2}$ per cent. of the last lines (Nos. 40 and 80) being found to be blank. If the blanks extend over the earlier lines, which are not marked on the W-Z forms, but not as far as the lines marked on the V form, this will lead to a slight deficiency in the proportion in the sample of entries in line 1 and the other lines marked on the W-Z forms. This disturbance, however, is only very small, but any tendency of the investigators to alter the order in which the names were entered so as to secure a suitable person for supplementary questioning could easily give rise to more serious biases.

The danger of this type of bias is always present in this method of sampling, and can only be overcome by the most rigorous training of observers, and the imposition of rules which determine uniquely the order in which names are entered on the list.

4.12 Frames from lists of households or dwellings

Lists of households or dwellings are frequently available from such sources as rating offices, electoral registers, etc. Frames based on such lists are in many ways preferable to frames based on lists of individuals. As already mentioned, in most surveys in which the information is collected by personal visit it is advantageous, and often essential, to collect information from all members of a household, in other words to use households as sampling units.

Frames consisting of lists of dwellings also have a much greater degree of permanence, being unaffected by movements of the population. Such frames, if complete at the time of their construction, will only become incomplete to the extent that there is new building, or changes in the use of existing

buildings. New building is necessarily a slow process, and the listing of new buildings usually presents no serious difficulties.

Lists of households can generally be utilized to give a frame of dwellings by taking as the sampling units the dwellings occupied by the households at the time the frame was constructed. Certain special precautions are required to ensure the inclusion of dwellings which were unoccupied at the time the list was prepared. In a town in which the dwellings are arranged in streets and in which the list is also arranged by streets this presents no particular difficulty. What has been called the *half-open interval* can be used. The procedure is as follows. When drawing the sample, the dwelling unit appearing next in the list to the selected unit is recorded, and the field investigator is instructed to see if there is any other unit on the ground between these two units, and if so to include that unit in the sample. Thus the field investigator might receive the instruction to survey No. 9 in a certain street, with No. 13 as the next recorded unit (odds only). If on visiting No. 9 he finds that No. 9A and No. 11 also exist, these are also surveyed. The even numbers between 9 and 13 are not included, since the instruction "odds only" indicates that they lie on the opposite side of the street. This procedure is clearly only possible if the list is arranged in an order which corresponds to some geographical pattern on the ground. If the list is not so arranged, incompleteness of the frame cannot be corrected by the use of the half-open interval or analogous procedure. In such cases the frame will have to be amended by other means, and complete rearrangement of the list in some geographical order may be necessary.

An example of the use of this type of frame is provided by some surveys carried out during the war in certain towns in the United Kingdom by the Ministry of Home Security, to investigate disturbances to the population on account of air raids. In the English towns electoral registers were used as frames, and in the Scottish towns rating lists were so used. The electoral registers consist of printed lists of voters arranged by streets in order of dwellings, all voters in one dwelling appearing together. Each dwelling therefore has as many entries as there are voters. Consequently, selection of entries in the list with equal probability will not give an equal probability of selection in the different dwellings. This could have been overcome by subdividing the list into dwellings, and basing the sampling on these dwellings.

As the surveys had to be conducted at considerable speed, delay in selection of the sample was avoided by the device of examining every x th entry, and including the dwelling in the sample if the entry referred to the first listed member in the dwelling. This introduces a certain additional discrepancy between the working sampling fraction, $1/x$, and the actual fraction of dwellings included in the sample, which introduces errors that are appreciable relative to the sampling errors for estimates of such quantities as numbers in the population obtained by multiplying the sample total by x . Such estimates were not the primary concern of these surveys, and consequently no adjustments were required. If necessary, errors arising from this cause could have been eliminated subsequently by ascertaining the ratio of the number of dwellings included

in the sample to the number in the whole register, and treating this ratio as the true sampling fraction.

In this survey the method of the half-open interval was used to deal with dwellings not included in the list, and was found to be quite satisfactory. Had there been new housing estates not covered by the register these would have had to be dealt with separately.

In certain towns the separate flats of blocks of flats were not listed, and therefore presented a difficulty, since the blocks constituted very large units whose chance inclusion or exclusion would have materially increased the sampling error. The existence of blocks of flats was, however, always apparent from the large number of voters appearing under the same address. The blocks were therefore listed, together with other large institutions, by preliminary inspection of the register, and every x th flat was selected by visit to all the blocks in turn.

4.13 Frames provided by town plans

Town maps and plans provide a useful frame for the sampling of dwellings in built-up areas. In some cases there may be detailed maps showing the location of all dwellings, but in many cases only street plans, not showing any great amount of detail, will be available.

Any town plan which gives an accurate representation of the streets will enable the town to be divided up into "blocks," *i.e.* areas bounded by streets. Such a plan will, therefore, provide a frame for area sampling in which the units are blocks. A sample of dwellings can then be obtained by including all the dwellings in the selected blocks. In general, however, the variability between block and block is likely to be large even after careful stratification, since there is often considerable local segregation of different classes of the population. Consequently, two-stage sampling is in general advantageous, blocks being taken as the first-stage units and dwellings as the second-stage units.

To obtain a two-stage sample in cases in which the map does not show the location of dwellings, it will be necessary to construct the second-stage frame for the blocks selected at the first stage by ground survey. This, however, is a much lighter task than the construction by ground survey of a frame for all dwellings in the city, and can frequently be done in the course of the survey itself.

In towns in which the natural blocks are of very unequal area, groupings of the smaller blocks or subdivision of the larger blocks should be performed, so as to reduce the within-strata inequalities in area. If little is known about the town, it may be advantageous to make a ground survey in order to demarcate and stratify the block units. It may even be advantageous to carry out a rough preliminary count, or make some other rough estimate of the number of dwellings in each block, as this will enable the subsequent selection of the first-stage units to be made from within strata with probabilities proportional to the

estimated numbers of dwellings. This was done in parts of the Greek population census described in Section 4.16. If such preliminary estimates are not available the best that can be done is to make a selection with probabilities proportional to area, but block areas are unlikely to be very closely correlated with the number of dwellings, even within strata. In either case the second-stage sampling fractions may be taken inversely proportional to the first-stage fractions, so as to give a constant overall sampling fraction.

Whether an elaborate procedure of this kind is needed depends not only on the accuracy required but also on whether estimates of total numbers are required from the survey. Since total numbers will be highly correlated with numbers of dwellings, prior supplementary information on these numbers for the different blocks, even if only rough, will be particularly effective in reducing the sampling variability of estimates of total numbers. They are not likely to have such large effects on estimates of the proportions of the population falling in various categories.

Sampling by streets is sometimes used instead of sampling by blocks. This is usually not so satisfactory as sampling by blocks, since each block represents a clearly defined area, whereas if a street is taken, there is often doubt as to exactly what is to be included and what excluded: alleyways and courtyards having entrances from more than one street, and not shown on the street map, for example, present considerable difficulty if the sampling is by streets.

Sampling by blocks is particularly valuable for surveys of towns in which all types of building have to be covered. Second-stage sampling of any or all of the different types of building can be adopted if required, by enumerating the different types for the selected blocks after the first-stage sample has been drawn.

4.14 Frames provided by maps of rural areas

The use of maps as a frame for the sampling of rural areas presents somewhat different problems from those encountered in the sampling of towns by the aid of town plans.

If accurate and detailed maps showing all or virtually all buildings are available, rectangular areas may be used as sampling units, the buildings falling in the selected areas being examined on the ground to see whether they are dwellings, with a further examination for unmapped dwellings.

Sampling with probability proportional to the apparent number of dwellings indicated by the map is possible, but would involve counting the dwellings in all the rectangular areas. Consequently it is better, if preliminary work on the maps of this magnitude appears to be worth while, to divide the map into areas containing approximately equal numbers of dwellings, using natural boundaries as far as possible and paying particular attention to stratification.

It may be noted here that the selection of a point at random on the map and selection of the dwelling unit nearest to this point for inclusion in the sample—a method which is sometimes used—is inadmissible, since a unit

which is widely separated from other units will have a much greater chance of selection than one which is close to other units.

With less detailed maps rectangular areas marked on the map will not be capable of being demarcated exactly on the ground. Natural features occurring on the maps must therefore be used as boundaries of the sampling units. This will necessarily give units of differing size. In particular, occasions will arise when it is impossible to subdivide a somewhat large area. In such cases, the area in question may be taken to represent two or more units. If any of these units are selected, a subdivision into two or more parts as alike as possible is made on the ground at the time of the survey, and the requisite number of parts selected by random choice.

In most countries, even in rural areas, there will be a number of villages of varying sizes, which are best dealt with separately by some form of stratification and two-stage sampling, since if these are included in the area sample a high degree of variability will be introduced. The use of a variable sampling fraction at the first stage, a larger proportion of the larger villages being selected, will be advantageous. A compensating reduction in the second-stage sampling fraction can be made if desired. The boundaries of all villages will require careful demarcation, as otherwise there will be ambiguity as to what should be included in the area sample.

4.15 Frames from lists of villages

In undeveloped areas the available maps are not likely to be of sufficient accuracy for area sampling. Where the population is concentrated in villages, these usually form the best first-stage sampling units. A list of villages will then serve as a suitable frame.

Even if the majority of the population is concentrated in villages there may be a residue located in the intervening countryside. If this residue owes allegiance to definite villages, the problem is relatively simple, since all that is required is the identification of the individuals belonging to the selected villages. This can normally be done by the head-men of these villages.

If no such association exists, some form of area sample of the intervening areas may be necessary. If rough maps are available, suitable areas may be demarcated by tracks, rivers, etc. If no maps are available, some form of line sampling may be possible in open areas.

If the country is not sufficiently open to be easily traversed, the construction of a rough map of the tracks may be necessary before any adequate sampling of the intervening areas can be carried out. It may be possible, however, to use these tracks without full mapping. Thus all tracks leading from villages selected for the sample may be traversed, and dwellings to which they give access included up to half-way to the next listed, but not necessarily selected, village. Such a method will only be effective with a relatively simple track system such as is met with in forest areas: intermediate junction points, for example, present special problems.

4.16 The 1946 population sample for Greece

The 1946 population sample for Greece provides a good example of the way in which sampling methods of the types discussed in the previous sections can be used to obtain speedy census data from a sample covering the whole population of a country. (Jessen *et al.*, 1947, C).

The sample was taken by the Allied Mission for Observing the Greek Election, as part of the investigation of the accuracy of the electoral lists. A population sample was required in order to test for omissions from the electoral lists, and the opportunity was therefore taken of securing more general census data. The corresponding test for duplications and redundancies in the lists was made by examining the lists themselves and investigating a sample of names drawn from them.

The frame for the first stage of the population sample was that given by the 1940 Population Census. This census gave returns for *koinotetes*, which are small communities or groups of villages, and *demoi*, which are towns and cities, usually with more than 10,000 population. Maps were available which showed the areas included in these *koinotetes* and *demoi*, and the names and location of all the populated centres. The *koinotetes* and *demoi* were used as sampling units at the first stage of the sampling. The units were stratified according to their population in 1940, and a variable sampling fraction was used. The actual scheme is shown in Table 4.16. Selection from within strata was systematic.

The sampling of the selected first-stage units was based on lists of households within the area. These lists were either based on existing lists checked and brought up to date, or were specially prepared to show the location of the households on a map.

For the sampling of towns an additional stage was used, a sample of blocks demarcated on an existing or a constructed street plan being first taken, with a further sample of houses from within the selected blocks. Sampling was sometimes with probability proportional to estimated numbers of households, these estimates being obtained by a rapid cruise of the whole area, and sometimes with equal probability.

The sampling fraction at the final stage was in all cases adjusted so as to give a constant overall sampling fraction. When blocks were sampled with probability proportional to estimated numbers of households, this required that the sampling fraction within the selected blocks should be taken as inversely proportional to the estimated number of households in the block. Thus the parish of Agios Panteleemous, which is given as an example, was initially subdivided into 98 blocks. Before sampling, some of the smaller blocks were combined so as to give 65 combined blocks.* The total of the estimated number of households was 966. It was decided to sample three blocks, which were selected systematically by taking a sampling interval of 322 ($= 966/3$) with a random starting point of 288, using sub-totals in the manner of

* This procedure is the same as that used in the sampling of Hertfordshire farms by parishes, Section 3.11.

Example 3.2.c. This gave combined blocks containing 23, 18, and 13 estimated households respectively. Since a sampling fraction of $1/100$ within the parish was required, the *estimated* number of households in the sample was $966/100 = 10$. This number was divided approximately equally between the three selected blocks (4, 3, 3), and the sampling intervals were calculated

TABLE 4.16—GREEK POPULATION CENSUS : SUMMARY OF THE SAMPLE DESIGN

Size-class code	Population in 1940	Assumed average population in 1940	Sampling ratios		Number of places	
			For selection of sample places	For selection of names and households within a sample place	In size-class	In sample
For <i>koinotetes</i>						
1	0- 499	350	1/100	1/5	2,147	20
2	500- 999	750	1/50	1/10	2,049	40
3	1,000-4,999	2,500	1/20	1/25	1,366	70
4	5,000 and over	7,000	1/5	1/100	54	10
				TOTALS	5,616	140
For <i>demoi</i>						
5	Under 25,000	17,000	1/2	1/250	52	26
6	25,000 and over	—	1/1	1/500	22	22
				TOTALS	74	48

by dividing the estimated numbers in the blocks by these numbers, *i.e.* the intervals were taken as $23/4 = 6$, etc. This procedure gives the required constancy in the overall probability of selection. The actual number of houses in the sample will of course differ from 10 if the estimated numbers are in error: it is the overall probabilities of selection, not the numbers of houses, that must be fixed.

The ratio method of estimation was adopted, using the 1940 population data as supplementary information. The actual method used was that appropriate to sampling without stratification with probability proportional to size of unit (Section 6.16), as this was considered to be the most accurate. There does, however, appear to be some danger of the introduction of bias by this method, and the unbiased method appropriate to a stratified sample with variable sampling fraction (Section 6.11) might have been preferable.

The survey was very successful and achieved high accuracy, the standard error of the estimate of the total population being estimated to be ± 2.1 per cent. The field work occupied 65 observer teams, each consisting of an observer, an interpreter and a driver, with a jeep, for three weeks. The entire sample and the computations were completed in 7 weeks.

4.17 Master samples

When a number of surveys covering the same population or aggregate of material are likely to be required, it is sometimes advantageous to construct a *master sample*, from which smaller samples can be drawn as required by means of a sub-sampling scheme.

The use of a master sample has a number of advantages. It enables a more accurate, complete and adequate frame to be constructed than could be justified if the frame were only required for a single survey. It simplifies the selection of samples, since in the sub-sampling only the material contained in the master sample has to be subjected to the selection process. It enables supplementary information to be obtained which is of value in improving the accuracy of the various surveys. And it enables surveys on the same material to be so planned that the same units are not selected an excessive number of times for different surveys—a matter of some importance when the information is obtained by response to questionnaires.

The most extensive and elaborate master sample so far constructed is the master sample for agriculture of the United States of America. The construction of this sample was undertaken by the Statistical Laboratory of Iowa State College, in co-operation with the Bureau of Agricultural Economics and the Bureau of the Census (King and Jessen, 1945, G).

The sampling units of this master sample consist of small areas covering the whole of the United States. The units have a mean area of about 2.5 square miles, but vary according to location and other circumstances, the mean area per state ranging from 0.71 square miles to 108 square miles. They were formed so as to contain on the average 4, 5 or 6 farms, depending on the part of the country. One-eighteenth of all the areas were selected for the master sample.

The whole of the land area of the United States was divided into three categories, called in the master sample "primary strata." These primary strata are (1) the incorporated stratum, (2) the unincorporated stratum, (3) the open-country stratum. The incorporated stratum consists of incorporated

cities and towns and unincorporated places regarded as "urban" by the Bureau of the Census. The unincorporated stratum consists of all named places outside the incorporated areas which have an estimated population of 100 or more, and all other areas which appear on the map and have a population density of 100 or more persons per square mile.

The incorporated areas were defined by the corporate boundaries, of which the location could be obtained. The unincorporated areas were demarcated on the maps so as to give areas as compact as possible, while including everything that did not appear to be open country. Subject to this, the boundaries were chosen so as to be easily identified on the ground. Aerial photographs were used in some cases for this work.

The general highway and transportation maps showed with varying degrees of accuracy the location of farms and other dwellings in the open-country areas and to some extent in the smaller unincorporated places, and these were therefore used to demarcate the actual sampling units of the open-country stratum. The procedure was as follows. The numbers of farms and non-farm units were first counted in what are termed "count units" from the map. A count unit consists of a unit defined by minor civil boundaries or natural boundaries, and in general included from 6 to 30 farms. These count units were numbered, and the number of farms and the total number of dwellings including farms were marked on the map. The number of sampling units into which each count unit was to be subdivided was also decided and noted on the map; in making this decision, consideration was given to the prevalence of natural boundaries, etc. The data for each count unit were then recorded on punched cards, and cumulative totals of the farm count, the dwelling count, and the number of sampling units, were tabulated. These cumulative totals were used to determine the count units which contained selected units, a random number between 1 and 18 being chosen as a starting-point, and the count unit containing every 18th sampling unit being selected thereafter. The count units containing selected sampling units were next subdivided on the map into the specified number of sampling units, the subdivisions being so chosen that they could be located on the ground. The units so demarcated were then numbered or counted systematically and the appropriate sampling units selected. Existing aerial photographs were used extensively for the demarcation of boundaries. In cases in which there were no suitable natural boundaries on the maps or photographs two or more units were amalgamated, subdivision and random selection being subsequently made on the ground if either unit was selected.

Somewhat different procedures, which need not be detailed here, were followed for the unincorporated and incorporated strata. For the incorporated stratum information was obtained from the Bureau of the Census on numbers of farms, etc.

In its final form the master sample will provide an adequate master sample of both farms and population, and also of the land area of the whole of the United States. Because the sampling units consist of areas, the frame will

remain complete and adequate whatever changes occur in the course of time. The supplementary information provided by the number of farms and number of dwelling units will naturally become progressively more inaccurate, but major changes are likely to take place only in limited areas, and the master sample will in the course of its use reveal the extent of these inaccuracies. There will, therefore, be no difficulty in revising the sample when it appears necessary for those areas of the country where extensive changes have occurred, and this will in no way invalidate the existing sample for the rest of the country.

It will be seen that the construction of a master sample of this type is a major undertaking, and it should not be assumed that a master sample of the same type is necessarily expedient in other countries in which the conditions are different. Thus in the United Kingdom the 6-inch Ordnance Survey maps provide an excellent frame for land area surveys, and the register of farms which is maintained for the collection of agricultural statistics provides a very complete frame of farms. If a master sample for agriculture is ever considered necessary, its construction could be based on this register and on the associated returns of farmers. The task of construction would therefore be very much simpler than would be the case if no such register existed. On the other hand, there is a need in the United Kingdom for an adequate master sample for localized population surveys. This problem is discussed in the next section.

4.18 Localized population surveys

As has already been indicated in Section 4.9, surveys are often required which will give reasonably accurate estimates for the country as a whole, but not for the separate small administrative districts. Such surveys have to be concentrated in a few localities, particularly if they are to be carried out by field investigators, since the amount of travelling would otherwise be excessive and supervision difficult. They may therefore be termed *localized* surveys. A multi-stage process must be used, the units at the first stage being administrative districts or similar areas of such size that each selected unit is capable of being covered by a single investigator or a small team of investigators.

The crux of the problem, therefore, consists of so planning the primary stage of the sampling process that the sampling error at this stage is not excessive. A secondary consideration, which must not be ignored, is that the within-strata comparisons should be sufficiently numerous to furnish a reasonable estimate of the sampling error at the first stage.

The use of stratification is obviously indicated. This stratification must in the first instance serve to differentiate between urban and rural areas. Consequently the country should be divided into large cities, into smaller urban areas, and into rural areas, in a manner somewhat similar to that employed in the master sample of the United States. The number of classes required will depend on the character of the towns and rural areas. A variable sampling fraction will be required in association with this stratification; for most surveys

it will probably be necessary to take all of the very large towns, but a proportion only of the intermediate towns, a smaller proportion of the smaller towns, and a still smaller proportion of the rural areas. Regional stratification of the smaller towns and rural areas may also be adopted as far as possible in parallel with this stratification.

These two types of stratification by themselves, however, are not likely to be entirely adequate for the urban areas, and some further form of stratification may be sought which will ensure (a) reasonably correct proportions of areas of different industrial types, and (b) reasonably correct proportions of the different social classes.

The methods by which it may be possible to ensure this will vary greatly according to the nature of the country, the type of primary unit that is adopted, and the amount of information that is available on these primary units. Administrative areas are usually most suitable from the point of view of the amount of readily available information, but they are not always ideal from the sampling point of view. As far as the United Kingdom is concerned, administrative areas appear to be the only possible type of area which can be used without a great deal of preliminary work. They will probably prove reasonably satisfactory if the boroughs and urban districts associated with the large towns are treated as parts of these towns, and sampled fairly intensively. Thus, for example, the sampling of the various parts of London and of its satellite suburban towns should be considered as a special problem separate from that of the sampling of the smaller towns in other parts of the country.

The second-stage sampling of the selected first-stage units is not likely to present any very serious problems. In the very large towns such as London, and in dispersed rural areas, two or more stages are likely to be required to avoid excessive travelling. Adjustment of the sampling fraction at the final stage to give equal overall sampling fractions is often advisable, since estimates can then be rapidly and simply obtained. Provision at the final stage for a proper rota of households to be included in the different samples, so as to avoid using the same household too frequently, is also of importance.

Much further research work remains to be done before it can be said with certainty whether a sample of this nature covering the United Kingdom is likely to be satisfactory for all purposes, or whether samples having a different structure will be required for different purposes. The importance of investigating the possibility of obtaining such a sample is clear. Without it, localized sociological and economic surveys of the general population cannot be carried out with any high, and at the same time ascertainable, degree of accuracy.

4.19 The U.S. series of employment estimates

An early example of a localized sample is that set up in the United States in 1939 to provide regular and speedy statistics on unemployment, employment

and the labour force (Frankel and Stock, 1942, F). The sample was modified and improved in 1943 (Eckler, 1945a, F).

In the original sample, counties were used as the first-stage sampling units. All the 3097 counties of the United States were classified and sampled as follows :—

	Total No. of counties	Percentage of population	No. of counties in sample
Cities	9	14	9
Urban	447	50	28
Rural	2641	36	27
	3097	100	64

The 9 city counties relate to the 5 largest cities ; all these 9 counties were included in the sample. The urban counties are those with 1930 populations of 45,000 and over. In the urban and rural classes a triple stratification, each of three strata, was adopted, the bases of the three stratifications being population, administrative areas, and percentage unemployed. Divisions between each of the main strata were so chosen that approximately equal numbers of counties fell in each main stratum. There were thus 27 sub-strata for both the urban and rural classes.* One county was selected from each of these sub-strata at random, with one exception where two counties were selected.

A further two-stage process was used to sample the urban and rural areas within the selected counties. The numbers of households to be selected from the various urban and rural areas were allocated on the basis of the census population figures for these areas. This led to the gradual development of a differential bias between urban and rural areas, owing to a drift of population away from the rural areas.

The results from within a single county were aggregated without any weighting. The aggregates were then weighted according to the population of the stratum from which they were obtained.

The within-county sample was changed every 4-6 months. This was a compromise between having a constant group of households, which would give most accurate estimates of monthly changes, and having new households on each occasion so as to avoid repeated visits to the same household. It introduced a certain discontinuity into the results, which has been avoided in the modified sample by using a proper system of partial replacement of the type described in Section 3.17.

In the modified sample, which included 68 first-stage units, allocation of households on the basis of population figures was abandoned. Instead, small areas were used as units at the second stage. This eliminates bias resulting from population drift.

* It will be noted that the numbers of the counties in the different sub-strata were not by any means equal.

Several other features were also introduced in order to improve the accuracy of the results. The stratification was more detailed, selection of the primary units was with probability proportional to their populations, and the ratios between the numbers of households having certain contrasting characteristics, *e.g.* farm and non-farm, were adjusted in each selected first-stage unit to agree with the corresponding ratios in the stratum to which the unit belonged. This last procedure is not entirely free from danger of bias. In a unit with a relatively small proportion of farm households, for example, those that do occur may be expected to be somewhat abnormal owing to proximity to non-farm areas, and such abnormal households will consequently receive excessive weight in the final results.

In both these samples only a single unit was selected from each of the first-stage strata. Although this unquestionably increases accuracy by permitting the use of smaller strata, it has the consequence that no fully valid estimate of error is available. The best estimate is that obtained by combining the strata in pairs, and this is likely to be somewhat of an overestimate.

4.20 Frames suitable for special classes of a human population

Surveys of special classes taken from the whole of a human population are often required. If a general frame covering the whole population is available, it can be used for a survey of a special class by selecting a sample from the whole population, and rejecting those members which do not fall in the required class. If the frame itself does not contain the necessary information, this will necessitate surveying all units of the sample in order to find out which individuals are to be retained and which rejected. If the required class is only a small fraction of the whole population there will be a large proportion of rejects, and a disproportionate amount of work is therefore required in such cases.

Consequently, if a frame covering only the required class or classes is available, this should be used in preference to a general frame. In surveys of the labour force, for example, it is often possible to utilize unemployment insurance registers and similar records. Such frames are often to a certain extent inadequate—all types of labour may not be included in an unemployment insurance scheme, for example—but their greater convenience frequently outweighs their defects. Occasionally it may be considered advisable to cover the excluded classes with lower accuracy by means of a general frame.

When no partial frame exists a survey undertaken for another purpose can sometimes be used to provide one. Thus in a recent survey of the aged carried out by the Nuffield Trust in certain towns of the United Kingdom, the records of an earlier survey by the Social Survey of the Central Office of Information covering all households were used to locate those households which contained aged people.

4.21 Frames suitable for the survey of economic institutions

Surveys of economic institutions may be divided into two general classes : those covering the whole or a large part of the commercial and industrial undertakings of a given town or country, and those covering a single type of undertaking or industry.

In surveys of the former class the use of frames constructed from maps or plans is often feasible : thus, in a general survey of factories of a given town a town plan may conveniently be used, with area sampling from this plan. Since, however, most commercial and industrial undertakings vary greatly in size, a variable sampling fraction is often required, a larger proportion of the large undertakings being selected. A map does not provide a suitable frame for this purpose. Even for general surveys, therefore, it is often advisable to use a special frame for the large undertakings, excluding these from the area sample, which is used to cover only the smaller undertakings. In a survey of the factories of a town, for example, there is usually no great difficulty in drawing up a list of the larger factories. If necessary a preliminary ground survey, with or without the aid of detailed maps, can be made.

If a particular type of undertaking or industry requires to be surveyed, the use of any form of area sampling will generally be unsatisfactory unless the units are small and widely dispersed, as, for example, occurs with retail shops. Even here shopping centres will require differentiation from other areas. In other cases a list of all the units of the given type of undertaking or industry will form a very much more suitable frame. In order that a variable sample fraction may be used it is important that such a list should contain some indication of the size of the units. If no satisfactory frame of this type exists it will often be worth while carrying out a complete census, simply for the purpose of constructing a frame and collecting a few basic facts about the given type of undertaking or industry. Such a census is usually more effective if it is on a compulsory basis.

When a frame provided by a complete census is required for repeated surveys, the problem of keeping it up-to-date must be considered. This is usually best effected by keeping a register of the undertakings concerned, and making a regulation that requires all changes to be reported. For the purposes of sampling, the most important type of change which requires to be recorded is that of new entries. Failure to report other changes will merely result in inaccuracies in the frame.

4.22 Market research and opinion surveys

Market research includes not only investigation into consumer reactions to goods and services and to advertising campaigns, but also investigations into consumer needs. In the case of consumer reactions information is mainly required on opinions, while in investigations of consumer needs factual information will also be required. Market research surveys can therefore be carried

out in the same manner as sociological surveys of the questionnaire type, using the methods which have already been described.

This is also true of other surveys of public opinion, but in many opinion surveys and also in certain types of investigation into consumer reactions, the requirements are somewhat different from those for sociological investigations. Speed is often essential, and changes in the percentage of individuals holding a given opinion are frequently of more interest than the absolute value of the percentage holding that opinion at any one time.

To meet these requirements, and to reduce costs to a minimum, what is known as the *quota method* of sampling has been developed. This method is a variant of purposive selection. Interviewers are given definite quotas of people in different social classes, of different age-groups, etc., and are instructed to obtain the requisite number of interviews in each quota. Additional instructions, which are designed to prevent excessively unrepresentative selection within the allotted quotas, may also be given on mode of contact, etc. The interviews themselves are sometimes carried out by house-to-house visits, sometimes by interviewing people in the streets and other public places, and occasionally even by telephone.

It is clear that, however accurately the quotas are fulfilled, such samples cannot be regarded as the equivalent of random samples. Consequently the danger of bias is always present, and the quota method must therefore be ruled out as a suitable method of investigation for precise enquiries in which unbiased results are required. Moreover, if there is a change in conditions, a quota sample which has previously adequately reproduced the characteristics of the population may cease to do so. Consequently, the fact that a quota system has consistently given reliable results over a period of years is no guarantee that it will also do so in the future.

The striking failure of public opinion polls to predict the results of the 1948 American Presidential Election—in contrast with previous successes—has been explained as being due to a shift of opinion during the closing weeks of the campaign, but it is possible that this failure was in part attributable to defects in the sampling procedure. It is generally agreed that the votes of organized labour influenced the outcome of the election to a much greater extent than usual, and it may well be that, in spite of the quota system, the samples were very deficient in factory workers and other trade union labour. The mere fact that a quota system is designed to give the correct proportion of workers, or even of different classes of workers, does not necessarily ensure that those included are representative, as regards the way they vote, of the workers as a whole. Consequently the results may be biased in elections in which the different types of worker vote very differently.

Apart from the problem of obtaining a representative sample, there is the inherent difficulty in opinion surveys that an individual's opinion on a given subject is frequently both ill-defined and liable to change. Moreover, on certain subjects the respondents may be unwilling to voice their true opinions. Opinions are also held with very different degrees of intensity, which there

is no easy way of measuring. Much of the information provided by public opinion polls is therefore of doubtful significance.

On the other hand, if used with skill the quota method may give sufficiently accurate results in simple enquiries where only general indications of the opinions held are required. If the samples are taken in the same manner on different occasions, and circumstances remain broadly the same, it may also provide a not-too-inaccurate measure of changes of opinion.

4.23 Frames for agricultural censuses and surveys

Agricultural censuses and surveys can be carried out in collaboration with the farmers, or in certain circumstances by direct observation without contacting the farmers. The latter method is in general only applicable to surveys of agricultural crops, and then only if all particulars required are ascertainable by inspection. For censuses and surveys of livestock the collaboration of the farmer is usually necessary, the essential difference being that livestock is mobile whereas crops are immobile. Collaboration is also obviously required if information relating to the farm as a whole is needed. In many countries contact with the farmer is advisable even for crop surveys, because exception may well be taken to the examination of a crop without the farmer's permission.

If a census or survey is to be conducted by contacting the farmer, the farm will usually form the sampling unit at some stage of the sampling process. A frame covering farms will therefore be required. Such frames are provided either by lists of farms, or by some form of area sampling which serves to locate the farmhouses. Frames based on maps, etc., which are suitable for the sampling of human populations in rural areas (Section 4.14) are equally suitable for the sampling of farms.

If contact with the farmer is not necessary maps can be used directly as a frame for crop surveys. Their use for this purpose is discussed in the next section. Even in this case, however, farms may well provide the best available frame.

In crop surveys the natural unit for many purposes is the field and not the farm. In cases where it appears advisable to obtain information for some only of the fields of a farm under a given crop, a further stage will have to be introduced into the sampling process. This inevitably results in a somewhat complicated sampling structure with different sampling fractions for the different parts of the sample, which in turn introduces complications into the analysis of the results, at least if unbiased estimates are required.

An example of this type of survey is provided by the Survey of Fertilizer Practice, carried out in various counties of England and Wales from 1942 onwards (Yates *et al.*, 1944, G). The objects of this survey are to determine the way in which farmers manure the different crops, and the relation of this manurial practice to the fertilizer requirements of the soil, in so far as these can be determined by the current methods of chemical soil analysis.

The method of selecting the samples is as follows. For each county a

systematic sample of farms is selected from the Ministry of Agriculture's addressograph list, maintained for the purpose of collecting the agricultural statistics on crop acreages and livestock. This list is arranged alphabetically by farmer's name and parish, and shows the total acreages (crops and grass) of each farm. A variable sampling fraction is used, with three size-groups, about 100 farms being selected from an average county. Larger samples are taken from counties which can be subdivided into districts containing different types of farming.

Each selected farm is visited by a field investigator, who is a member of the Provincial Advisory Staff. All the fields of the farm are listed in consultation with the farmer according to their crops, and also according to whether they have been recently ploughed out from grass (*new* and *old* arable). In the earlier surveys one field of each crop was selected at random from all the old-arable fields, and similarly for all the new-arable fields. One permanent grass field was also selected. In the later surveys one field in three of each crop has been selected from each of these categories. From each group of selected fields one old-arable, one new-arable and one permanent grass field is selected at random, and soil samples taken for chemical analysis.

For the selected fields information is obtained from the farmer on the cropping over the previous four years, and the amounts and chemical composition of the fertilizers, farmyard manure and lime applied in each year of this period. In some of the later surveys only a single year has been covered. When necessary the fertilizer merchants are consulted in order to obtain information on the chemical composition of the fertilizers.

The methods of analysis adopted in this survey are illustrated in Example 6.19.

4.24 Use of maps as frames in agricultural surveys

If accurate large-scale maps showing the field boundaries are available, the point method of sampling is very suitable for crop surveys in which contact with the farmer is not necessary. The fields will then act as sampling units, and selection will be with probability proportional to size. Provided the whole of a selected field is under a single crop, all that is necessary for acreage estimates is to ascertain the crop, no determination of area being required (Section 3.9). If more than one crop is being grown on a selected field, the proportions of the area under the different crops must be determined, but eye estimates will usually be adequate for this purpose.

In this type of work two-stage sampling will often be advisable in order to save travelling, and also to avoid having to handle an excessive number of maps. Thus in the United Kingdom the 6-inch Ordnance Survey quarter-sheets (3 miles $\times$ 2 miles) might provide suitable first-stage units, a fairly dense grid of points being taken over the selected sheets.

If selection with equal probability of irregularly-shaped areas such as fields is required, these areas must each be defined by a single point, such

as the most northerly point of the area. The map is then divided into sampling units consisting of rectangular areas, a number of which are selected with equal probability, the fields with defining points in the selected areas being included in the sample. Only the selected rectangular areas need be demarcated. If more convenient, circular areas whose centres are located at random (or systematically) may be used in preference to rectangular areas. Rectangular areas have the formal advantage that the whole of the area is included once and once only in the aggregate of sampling units, but this is not of great practical importance.

It should also be noted that with this method of selection the sampling units consist of groups of fields whose defining points are included in a single rectangular area, and not the fields themselves. Rectangular areas which contain no defining points must be counted as units of zero area. It can be shown that when the rectangular areas are small and mostly contain one or no defining point, the sampling errors of estimates of crop acreages are greater with this method of sampling than with the point method. The point method is therefore preferable under these circumstances.

On the other hand, if the rectangular areas are large relative to the sampled fields—as will be the case, for example, if whole sheets of a map are surveyed—the use of defining points in this manner saves splitting fields which are cut by the map boundaries. Some slight additional variance will be introduced unless the total area of all fields is determined and used as supplementary information.

Maps have been extensively used as frames for the estimation of the acreages of crops in surveys conducted by the Calcutta Institute of Statistics (Mahalanobis, 1944, A ; 1946, A ; 1940, H ; 1945, H ; 1946, H). The method followed is to demarcate square areas located at random on the maps, and to survey all fields covered in whole or in part by these areas. The areas of the whole and part fields are determined from the maps by measurement. This measurement of areas and their subsequent summation might be avoided by the use of point sampling: if each square area were replaced by a square pattern of 9 or 16 points, for example, it would appear that the loss of accuracy would be small (see Example 8.11.c).

4.25 The 1942 Census of Woodlands

The 1942 Census of Woodlands covering England and Wales provides an example of the use of maps as a frame. The object of the survey was primarily to determine the volumes of standing timber of various types in the country, and their broad regional location, in order to estimate the amount of available home-grown timber and to plan its utilization.

The sample was initially planned to be taken in two parts, each consisting of 5 per cent. of the total land area. The sampling units were 6-inch Ordnance Survey quarter-sheets (3×2 miles), systematically located on two interpenetrating 12×10 -mile rectangular grids, one for each part. All areas of

woodland of over 5 acres on the selected quarter-sheets, and 1 in 5 of the areas under 5 acres, were surveyed. Areas of woodland cut by the boundary of the map were surveyed if their southernmost point was included in the selected map, areas subdivided by rides marked on the map being treated as separate areas for this purpose. The land areas covered by the selected maps were also inspected in the course of the survey to determine any new plantings since the map was last revised, thus correcting for any incompleteness in the frame.

The woodland areas were divided by inspection on the ground into "stands," each of which represented a homogeneous area of woodland. The boundaries of these stands were demarcated on the maps so that their areas could be determined, and a representative plot was chosen from each stand on which all or a sample of the trees were measured. Representative plots were used instead of random plots because reasonably accurate volume figures for the individual stands were required. Control of the bias introduced by the use of representative plots was effected by determining the quantities of converted timber actually obtained from surveyed stands felled in the course of ordinary forestry operations. This procedure served also as a check against any errors in the assumed wastages on conversion. A further control by the measurement of randomly selected plots on a sub-sample of stands was also planned, but was not in fact carried out.

The total area of woodland was determined independently from the areas coloured green on the 1-inch Ordnance Survey sheets (see Example 7.18). Errors in the 1-inch sheets were allowed for by comparing the selected 6-inch sheets with the 1-inch sheets after survey. The final estimates of volume were calculated from the volumes per acre determined from the surveyed areas and the total areas determined as above.

A first estimate of volumes was required within six months of the decision to undertake the survey, and it was thought that with the teams available the first part of the survey could be completed and the estimates prepared within this time. Before field work commenced, however, it became apparent that the original programme could not be adhered to. Each selected quarter-sheet was therefore roughly divided into two halves as similar as possible, and one half of each sheet was selected at random for the first part of the survey, giving a $2\frac{1}{2}$ per cent. sample of all woodlands in the country. Subsequent calculations of the sampling errors showed that this $2\frac{1}{2}$ per cent. sample was quite adequate for the determination of general policy, which was the first objective of the survey.

The survey was then completed in two further parts, first the remaining halves of the first set of quarter-sheets, and secondly the other set of quarter-sheets. The three parts therefore consisted of $2\frac{1}{2}$ per cent., $2\frac{1}{2}$ per cent. and 5 per cent. respectively of all the woodlands of the country. In addition certain heavily wooded areas were completely surveyed.

The history of this survey demonstrates the extreme flexibility of sampling surveys, and the way in which they can be made to yield preliminary results

of ascertainable reliability. By the procedure of first surveying a properly selected quarter of the whole sample, it was possible to obtain the preliminary estimates in the required time in spite of unexpected delays in the commencement of the survey.

4.26 Frames for undeveloped areas

If no accurate maps are available, exact location of previously demarcated small sample areas on the ground will be impossible. Alternative methods must therefore be employed.

For completely undeveloped areas such as natural forests the line method of sampling is very suitable, provided the terrain and vegetation is such that the lines can be followed on given compass bearings without an undue amount of deviation. Distances along the lines can be determined by some simple measuring device such as a rope, or even by pacing. Where volume measurements are required small areas can be demarcated at given distances along each line.

Some frame for the location of the lines is necessary. This can often be provided by existing mapped roads or other tracks, but it is by no means impossible to construct a secondary frame as the survey proceeds by the use of cross traverses, using any available tie-in points. Except where maps are to be constructed, no great accuracy in the location of the lines is required, since it is only necessary that they be located in an unbiased manner with a density which is the same for the different parts of the area, or, if not the same, is determinable.

In areas in which a line on a fixed bearing cannot be followed, any attempt at complete and unbiased coverage must necessarily be very expensive. Often, however, a sufficiently unbiased sample of natural vegetation will be obtained by traversing existing tracks and taking sample areas at suitable intervals by offsets at right angles to the tracks. If a map of these tracks is not previously available it may be worth constructing one by rope and sound or similar rough surveying technique.

Crop surveys in partially developed areas without adequate maps present somewhat different problems. If the cultivated areas are located in the neighbourhood of villages, a two-stage sampling process will probably be required, a sample of villages being taken at the first stage. Since the total area of cultivated land associated with a village is likely to be closely correlated with the population figures, these (if known) should be treated as supplementary information. If not known the feasibility of making a simultaneous population census should be considered, since information on cultivated areas will be of more value if it can be related to population figures. In this case the sampling may well be two-phase, a larger sample being taken for the determination of population.

The survey of the cultivated areas associated with the selected villages will require the construction of second-stage frames. If the line or point

method of sampling is practicable this is likely to be the simplest method of dealing with compact areas of cultivation. Outlying fields will in this case have to be enumerated and sampled separately.

In many cases enumeration of all fields will be the only practicable method. The preparation of a sketch map will then be advisable. A certain percentage of the enumerated fields can be measured for area, and the crops determined if this has not been possible at the mapping stage. If the cropping is known, stratification by crop should be made before the selection of the sample for area measurements. A frame of this kind may remain serviceable, with some revision, over a number of years. It will also serve to locate the samples required in a crop estimation scheme.

4.27 Use of aerial survey photographs

When no maps are available the possibility of using aerial survey photographs as a frame for agricultural and land utilization surveys should be borne in mind. Although it is unlikely to be practicable to make an aerial survey simply for the purpose of providing a frame for sample surveys, it is often possible to utilize a survey that has been undertaken or is contemplated for other purposes.

Any aerial photographs covering the area are likely to provide an adequate frame, though the use of aerial survey photographs even for a frame is not as simple as it appears at first sight. The mere handling of the photographs covering any large area is a somewhat difficult task which demands an adequate and properly trained office staff. Moreover, aerial photographs are subject to variations of scale (and also distortion) due to tilt and changes of altitude of the aircraft. The stated scale is therefore not always correct, and the scale sometimes exhibits disconcerting variations even over different parts of the same mosaic. The precaution should therefore be taken of checking the scale by means of measurements on the ground in a sufficient number of instances to make certain that no important source of error is introduced.

Various methods of sampling can be used in conjunction with aerial photographs. If crop acreages have to be determined, point sampling is suitable. After the points have been marked on the photographs the fields in which these points fall must be identified on the ground and the crops growing on them recorded. In order to avoid excessive travelling it will almost certainly be worth using a two-stage process, the units at the first stage being rectangular areas which can be demarcated on the photographs, with a number of points taken within each of the selected areas.

If line sampling is required, the lines can first be demarcated on the photographs, and subsequently surveyed on the ground. In certain circumstances it may be possible to make the intercept measurements on the photographs, using the ground survey merely to determine the characteristics of the various intercepts.

If areas such as fields, the boundaries of which are recognizable on the photographs, are to constitute the sampling units, they may be selected with probabilities proportional to their sizes by point sampling. If there are likely to be ambiguities in the definition of the boundaries the units should be demarcated before selection.

If natural units such as farmhouses which depend on point locations are to be selected, small rectangular or circular areas may be used as sampling units in the same manner as in selection from a map.

In certain cases aerial photographs may provide the necessary information without any ground survey work. It is usually possible, for example, to recognize cultivated areas on the photographs, and the total cultivated area may consequently be determined directly from the photographs. In certain cases it may even be possible to differentiate between the different crops. In these cases the total cultivated area and the proportions of the area under the different crops can be determined by sampling of the photographs, point or line sampling being used as convenient. If desired, adjustments for variations in scale can be made by varying the spacing of the points or lines.

In some cases the differentiation between the different crops on the photographs may be only partial, or subject to error. In such cases a sub-sample of the points classified on the photographs can be re-classified by ground survey. The information provided by the photographic classification will then serve as supplementary information. By this procedure the amount of ground survey necessary may be very considerably reduced. The examination of stereo-pairs may be a considerable aid to the classification of certain types of area, particularly forest areas.

If an aerial survey is specially undertaken for the purpose of a sample census or survey, it is possible to reduce the amount of photography by taking parallel strips of photographs separated by unphotographed areas, but aerial photographs taken in this manner will not be of much use for mapping purposes. If no map frame is available, a few cross-strips will have to be taken to provide links between the separate strips. Too much reliance must not be placed on the accuracy of the location of the strips unless special navigational aids are installed.

4.28 Crop estimation

The total yield of a crop can be regarded as the product of its acreage and the mean yield per acre. These two quantities may therefore be estimated separately. Estimates and forecasts of the mean yield per acre must of course be related to the conventions adopted in the estimation of acreage, particularly with regard to areas on which the crop has failed or been abandoned.

The estimation of acreage has already been discussed in the preceding sections, and in this section we shall therefore mainly be concerned with the problem of the estimation of the mean yield per acre.

There are a number of ways in which estimates of the mean yield per acre

of a crop, or the total yield, may be obtained. These may be broadly classified as follows :—

- (1) Reports from crop-reporters, who, at or subsequent to harvest, make returns to a central authority of their estimates of the average yields of the crop in their own districts, these estimates being based in the main on general impressions, discussions with farmers, etc.
- (2) The harvesting of small sample areas of the crop immediately prior to the main harvest.
- (3) Eye estimates of the yields of a sample of fields, with subsequent calibration of these eye estimates by comparison with the actual yields of some at least of the sample fields.
- (4) Co-operation with the farmers at harvest time so that accurate yield figures may be obtained from a sample of fields as they are harvested.
- (5) Returns by farmers of the yields of their crops.
- (6) Market returns, export statistics, etc.

If necessary, Methods (2) and (3) may be combined in a two-phase sampling scheme, eye estimates being taken from a comparatively large sample of fields, with crop-cutting samples from a smaller sub-sample of these fields.

These various methods all have their advantages and disadvantages. Method (1), that of crop-reporters, is the one commonly adopted by countries with long-established and stable systems of agriculture. Its success depends on the ability of the individual crop-reporters to make accurate and unbiased estimates of the average yields of their districts. The method is not objective, and no assessment of its accuracy can be made unless independent estimates, provided by some other method of known or ascertainable accuracy, are available for comparison. Doubt is often cast on estimates provided by the method because of disagreement with market returns, etc., and their lack of objectivity makes it impossible to say which set of estimates is at fault.

Even if crop-reporters are reasonably accurate on the average over a run of years, estimates for particular years or particular districts may be subject to considerable errors. There seems to be a general tendency, for instance, to underestimate yields in good years and overestimate them in bad years. The accuracy attained may also be very different for the different crops. Moreover, spurious long-term trends may be introduced through gradual changes in the standards of the reporters, and this considerably reduces the value of the estimates as a measure of the improvement or deterioration of the agriculture of a country. Any sudden change in an agricultural system, such as the introduction of new varieties, or the bringing into cultivation of new land, may introduce disturbances into previously satisfactory estimates.

Method (2), the harvesting of small sample areas, is theoretically capable of providing a completely objective estimate of the mean yield per acre of the standing crop at harvest time; it will not, of itself, provide any estimate of the losses at or subsequent to harvest. In practice, however, serious bias

may arise in a number of ways if proper precautions are not taken. These sources of bias, and the practical details of the method, are discussed further in the next section.

Method (3), that of eye estimates, has the advantage that on certain types of crop such estimates can be relatively rapidly made, and consequently a larger sample of fields can be visited in a given time. The difficulty of having to transport and thresh a large number of samples, which often arises with Method (2) is also avoided. Some results of a trial of this method on wheat are given in Example 6.15. The method is not suitable for root crops such as sugar beet and potatoes, since it is difficult to judge the yields from inspection of the tops. In such crops, however, there are no transport and threshing problems, since the samples can be weighed in the field.

If calibration of the eye estimates from the farmers' yields is not practicable, or if the calibration is found to vary substantially from year to year, a few specially-trained field workers can be used to take crop-cutting samples in order to calibrate the eye estimates of each investigator at the time of harvest.

Methods (4) and (5) require the co-operation of the farmer. Method (4) differs from Method (5) in that in Method (4) the harvesting is done in the presence of an investigator, and if necessary with assistance, such as the provision of a threshing machine, whereas in Method (5) reliance is placed entirely on the farmer to provide accurate yield figures. Owing to delays of threshing, etc., Method (5) is not likely to provide estimates till some time after harvest.

Estimates from market returns, export statistics, etc. (Method 6) provide a useful basis for comparison with estimates by other methods, but such returns will only exceptionally give an accurate estimate of the actual yields, since the amount of the crop passing through the market is likely to vary very considerably in different circumstances.

In Methods (2) and (3), which require field investigators, the question must be considered whether the survey should cover the whole of the country or whether it should be confined to certain districts only, using a two-stage sampling process. If only an estimate of the yield of the whole country or of large districts is required, comparatively few fields will need to be sampled, and a single-stage process for the selection of fields will result in a very dispersed sample. A two-stage process will avoid this difficulty at the cost of introducing a between-districts component of variation into the sampling error.

Finally, it should be emphasized that crop estimation, though theoretically simple, presents many practical difficulties. The introduction of a satisfactory scheme where none exists, or the provision of objective estimates to check existing subjective estimates, requires continuous work over a number of years by a properly established team of workers. Except for preliminary investigations, crop-estimation projects should therefore not be undertaken unless continuity can be maintained. Nor should an existing method of estimation be abandoned or disturbed until a better alternative has been evolved and kept in operation for some time. If the old and new methods are run in parallel for a number

of years it will be possible to assess the reliability of the old method and its degree of bias, a task which will be quite impossible if it is discontinued before adequate comparative data have been obtained.

4.29 Estimation of yield by the harvesting of sample areas

As mentioned in the previous section, the estimation of the mean yield per acre of an agricultural crop by the harvesting of small sample areas presents many practical difficulties, and the results may be biased in a number of ways if the proper precautions are not taken.

Errors can occur through faulty selection of the fields, through faulty sampling of the selected fields, through failure to take samples from the fields at dates sufficiently near harvest, or through failure to sample some of the selected fields owing to their having been harvested before they were visited.

If rigorous means of selection are employed there is no reason why the selection of the fields should be faulty. If, however, the cruise method is used, fields being taken at equal distances along a given route in the manner described in Section 3.15, the estimate will almost certainly be appreciably biased, though this bias may be reasonably constant from year to year if the same route is followed each year. On the other hand, the use of the cruise method overcomes the difficulty of ensuring that the visits to the fields are made sufficiently near harvest, and also eliminates the risk of missing fields through their having already been harvested. All that is necessary is to traverse the route at sufficiently close intervals of time, stopping the car at each sample point and examining the crop to see if it has reached a sufficiently mature stage for a sample to be taken.

An alternative procedure which is sometimes used is to follow the prescribed route and take a sample from all or a given fraction of the fields that are actually being harvested. This, however, may introduce an additional component of bias, since, unless special precautions are taken, limitations of time will result in the inclusion of a greater proportion of the fields which are harvested very early or very late.

With crops that do not have to be fully mature at harvest, *e.g.* potatoes, samples will normally have to be taken somewhat before maturity, unless information is available from the farmers as to when they intend to lift. With such crops, however, it is usually possible to estimate the amount of growth between the time of taking the sample and date of harvest; this latter date can then be determined by a subsequent visit. In the potato crop, for example, investigation has shown that the weight of tops provides a fair indication of the amount of further growth that may be expected.

The cruise method of sampling, therefore, provides a method of crop estimation which, though theoretically more liable to bias than a proper random selection of fields, may in practice give more satisfactory results, particularly in the estimation of yields per acre. It is also likely to be considerably more

economical in travel. Which method is most suitable will depend largely on local conditions, and must be the subject of local investigation.

Bias in the estimation of the yields of the actual fields can arise from improper location of the samples and from cutting a larger area of the crop than the true unit area. An example of such bias has already been given in Section 2.5.

Edge effects are also liable to give rise to bias, since in an irregularly shaped field it is impossible without a great deal of labour to locate samples properly at random over the whole of the area. The method described in Example 3.2.b is clearly impracticable, and no simple method of traversing the field has been devised which will give equal probability of selection over the whole field. In practice, however, a systematic method of selecting the sample is quite adequate. The important thing is to see that the location of the sample units is as objective as possible.

The determination of the bias arising from headlands, lower yields at the edges of the field, and errors in estimation of the area of the field—the U.K. Ordnance Survey areas, for example, include farm roads, hedges and ditches—can be made if required by more rigorous supplementary observations on a small percentage of the fields. Often, however, the separate determination of these components of bias is of no great practical interest, since the losses at and after harvest will also affect the total amount of the crop that is finally available for consumption, and the total bias is best determined by comparison with the farmers' reported yields on a sub-sample of the fields, or by determining these yields in co-operation with the farmer.

Two methods of locating the sample units have been found convenient in practice in this country. The first is to traverse the field diagonally from corner to corner, using one or both diagonals, and locating samples at equal intervals along these diagonal lines. The interval required can be calculated by pacing the diagonal or making an eye estimate of the number of paces. Errors in the eye estimates are of little consequence, since the exact number of sampling units is immaterial. Alternatively, if the crop is in rows the field can be traversed along the rows. The length of one end is paced from corner *A* to corner *B*, the field being entered at a distance of one-quarter of this length from corner *B*. It is then traversed along this row to the other end of the field, and a return row is selected by the same procedure. A suitable number of sampling units is taken at each traverse in the same manner as in the case of a diagonal traverse. In this method of sampling it is advisable to step laterally across a given small number of rows after each sampling unit has been taken, since a given row may fall wholly on a particularly good or bad strip of the field, *e.g.* on a ploughman's "land."

Whatever the exact method of traversing the field, it is of the utmost importance that the location of the rows and of the sampling units should be made without inspection of the crop in the neighbourhood. This can be done quite effectively by counting paces and digging in the heel when the

requisite number of paces have been taken, but the field workers must be thoroughly trained in this procedure.

A good deal of work has been done on the most suitable size and shape of the sampling units. In this country experimental tests have shown that 4-6 contiguous quarter-metre row lengths are suitable for cereal crops, with 6-10 units per field; for potatoes 4 units each of 6 ft. of row are now being tested on a large scale. Mahalanobis (1946, A), working in India, has used three or four concentric circles of 2-8 ft. radius, each annular ring being harvested separately so as to provide a control of field workers—with ordinary workers a bias is regularly found with the smallest circle—and this gives a check that the samples have really been taken in the prescribed manner.

The best size and shape of the sampling units depends very much on the nature of the crop and local conditions, such as type of field worker, whether the crop is sown or planted in rows or broadcast, variability within fields, available equipment for threshing and transport, etc. Local investigation should therefore always be undertaken if any extensive work is contemplated. On the other hand, it should be recognized that the sampling error of individual fields is usually small relative to the variation from field to field, and consequently the introduction of a crop-estimation scheme need not await the results of such investigation; any reasonably efficient method will give satisfactory results provided bias is avoided. (See Section 8.12.)

4.30 Crop forecasting

The term *forecast* is here used to denote an estimate of the yield of a crop furnished at some date well before harvest. The term is sometimes used to indicate estimates made by crop reporters at or even shortly after harvest, since such estimates are usually subject to later revision in the light of information received from farmers. Such estimates, however, are better termed *preliminary* estimates, in contrast to the revised or *final* estimates.

There is some confusion also between forecasts and estimates of acreages, forecasts of mean yields per acre, and forecasts of total yields. Once the crop is sown the determination of the acreage, apart from crop failures, is a matter of estimation and not of forecasting, and any forecast of the total yield is usually best presented in the form of an estimate of the total acreage and a forecast of the mean yield per acre.

There are three main methods of crop forecasting. Forecasts can be provided by crop reporters, they can be based on meteorological data such as rainfall obtained prior to the date of the forecast, or they can be based on observations and physical measurements of the growing crop, alone or in conjunction with meteorological data.

Meteorological data do not directly provide forecasts of the yields. If they are to be used as a basis of such forecasts, reliable data on both the yields and the meteorological events must be collected over a number of years, and the crop-weather relations evaluated. The same is true if observations and

measurements on the growing crop are to be used. The evaluation of these relations requires the application of the method of statistical analysis known as "regression analysis." We shall not describe this method here, but it may be well to emphasize that its application is not entirely simple, and the advice of a mathematical statistician experienced in this type of work should therefore be sought.

It must not be assumed that it will be possible to evolve a prediction formula which will give satisfactory results, even if accurate and extensive data are available. In the first place, the yield of a given crop is influenced by meteorological and other events up to and sometimes after harvest, and this may introduce too great a degree of uncertainty into yields predicted some months before harvest to make the prediction of any value. In the second place, although meteorological factors undoubtedly account for a good deal of the variation in crop yields, they are not by any means the only factors. Changes in variety, insect pests, plant diseases, exhaustion of the fertility of the soil, changes in the type of land under crop, changes in the amount of fertilizers, and many other factors may also exert a major influence. Thirdly, meteorological effects are often somewhat complex, and it may therefore be impossible to determine them from a set of data extending over a limited number of years; owing to the similarity of weather conditions over large areas, data from a number of districts in any one year are only a partial substitute for data extending over a number of years.

One advantage of using measurements of crop growth instead of relying wholly on meteorological observations is that the crop is thereby used as its own integrator of meteorological and other effects up to the time of the measurements. Frost and flood damage, for instance, are clearly better assessed, once they have occurred, by survey of the crop than by examination of meteorological records. The selection of the particular types of observations and measurements which are likely to give an adequate basis for forecasting is a problem on which further scientific research is required, particularly in the case of grain and other seed crops. In the case of root crops investigation has already shown that the amount of growth made by the tubers or roots, coupled with some measure of the extent to which the plant is still growing, *e.g.* weight of tops, are likely to give satisfactory results.

Since the evolution of a satisfactory method of crop forecasting demands a knowledge of the actual yields over a period of years, an investigation of suitable methods can be combined with an objective crop-estimation scheme. Once the observations and physical measurements which are likely to give useful information have been decided, all that is necessary is to take these measurements on a sub-sample of the fields which will subsequently be selected for sampling at harvest. In the initial stages it may be better to carry out the observations on a special sample of fields, rather than on the more scattered sample which will be suitable for crop estimation proper. More intensive investigations can also be carried out on experimental plots on which different varieties are sown, and which are subject to different cultural treatments and

sowing dates. Experimental plots by themselves, however, are not likely to provide all the information required for the evolution of a suitable forecasting scheme, since the variation from field to field in a given district is often quite large, and the inclusion of a number of fields in the district usually gives a much more adequate representation of the average meteorological effects in that district than will a single field.

If observations and measurements are to be made on the growing crop, a sampling scheme will have to be devised in order that single plants or small areas may be selected for measurement. A method of selecting wheat shoots for height measurements, for example, is described in Section 2.4. The principles to be followed in the location of the sampling units in the fields or experimental plots are similar to those which operate in the selection of samples for yield estimates.

4.31 Determination of the size of sample when the sample is fully random

As has been indicated in Chapter 3, the size of sample required to achieve a given accuracy depends on the variability of the material and the extent to which it is possible to eliminate the different components of this variability from the sampling error. In this and the following section we will describe the procedure which is appropriate for determining the size of a random sample, and indicate the general relationship between the errors of a random sample and other types of sample. Detailed consideration of the more involved types of sampling must be deferred till Chapter 8, where the comparative accuracy of the various types of sampling is discussed.

In the discussion of sample size we shall require the concept of *standard error*. As already explained in Section 3.7, the sampling standard error of an estimate is a measure of the average magnitude of the random sampling error to be expected in that estimate. It also provides an indication of the frequency with which errors of various magnitudes may be expected to occur (Section 7.3). In rough general terms, one-third of the actual sampling errors will be greater than the standard error, and one-twentieth will be greater than twice the standard error.

In the case of a fully random sample from a large population the formula for the standard error of the estimate of the proportion of *units* of a given type, *i.e.* having a given *attribute*, is very simple. If **p** is the proportion of units of the given type in the whole population, and **q** = 1 - **p** is the proportion not of the given type, the standard error of the proportion of units of the given type in a random sample of *n* units (which provides an estimate **p** of **p**) is given by

$$\text{standard error of } \mathbf{p} = \sqrt{\frac{\mathbf{p} \mathbf{q}}{n}}$$

The formula holds unchanged if the proportions are replaced by percentages. Thus

$$\text{standard error of } (p\%) = \sqrt{\frac{(p\%)(q\%)}{n}}$$

The full line of Fig. 7.4 provides a graphical representation of the standard errors given by this formula. If 20 per cent. of the units of the population are of the given type, for example, the standard error of the percentage of units in a sample of 100 is

$$\sqrt{\frac{20 \times 80}{100}} = 4.0$$

which is the value given by the full line. Thus estimates in the range 20 ± 4 per cent., i.e. between 16 and 24 per cent., will be obtained in two-thirds of all samples of 100 units, and estimates in the range $20 \pm (2 \times 4)$ per cent., i.e. between 12 and 28 per cent., will be obtained in nineteen-twentieths of all samples. If a sample of 1000 is taken, the standard error will be 1.26, and estimates between 18.7 and 21.3 per cent. will be obtained in two-thirds of the samples (Section 7.4).

When dealing with estimates of quantities such as means and totals it is best for our present purpose to work in terms of the percentage standard errors. The *percentage standard error* of the estimate of a quantity is the standard error of the estimate expressed as a percentage of the true value of the quantity.

The percentage standard error of the estimate of the total number of units having the given attribute in the population is the same as the *percentage standard error of p*. From the above formula we see that this percentage standard error is given by

$$\text{percentage standard error} = 100 \sqrt{\frac{q}{np}} = 100 \sqrt{\frac{(q\%)}{n(p\%)}}$$

The percentage standard errors given by this formula are shown by the dotted line in Fig. 7.4.

In a population in which 20 per cent. of the units possess the given attribute, for instance, the percentage standard error with a sample of 100 is*

$$100 \sqrt{\frac{80}{100 \times 20}} = 20 \text{ per cent.}$$

If there are 10,000 units in the whole population, 2000 will possess the given attribute. The standard error of an estimate of this number from a sample of 100 will therefore be $2000 \times 20/100 = 400$, i.e. in two-thirds of the samples estimates between 1600 and 2400 will be obtained.

* This result can also be obtained directly from the actual standard error of the percentage, which is $100 \times 4/20$ i.e. 20 per cent.

The formula may be re-written so as to give the number n required for the sample when the required standard error of p or the required percentage standard error is known. We have

$$n = \frac{pq}{(\text{required standard error of } p)^2} \\ = \frac{10,000 q}{p (\text{required percentage standard error})^2}$$

The proportions p , q and p may be replaced by the corresponding percentages without change.

If, for example, we are sampling a population in which it is believed that about 20 per cent. of the units are of a given type, and it is required to determine this percentage with a standard error of 1 per cent. (*i.e.* a percentage standard error of 5 per cent.), we shall require to take a sample with a number of units given by

$$n = \frac{20 \times 80}{1^2} = \frac{10,000 \times 80}{20 \times 5^2} = 1600$$

These formulæ hold only for a random sample in which the sampling units are the units for which the proportion having a given attribute requires to be estimated. In sampling a human population, for example, the sampling unit may be the household and not the individual. In this case the standard errors of proportions of individuals determined from the sample will be larger—often substantially larger—than those given by putting n equal to the number of individuals in the above formulæ (Section 7.7 and Example 7.8.b).

The corresponding formulæ for a quantitative character are equally simple. We then have for the estimate of the mean or total from a random sample of n from a large population

$$\text{percentage standard error} = \frac{\text{percentage standard deviation of a unit}}{\sqrt{n}} \\ n = \frac{(\text{percentage standard deviation of a unit})^2}{(\text{required percentage standard error})^2}$$

In order to use these formulæ we require an estimate s of the *standard deviation* (or standard error) of a unit in percentage terms. The standard deviation of a unit is the measure of the variability of the units amongst themselves, and can be determined from a frequency distribution of the unit values. Table 7.1 provides an example of such a frequency distribution, and Example 7.1.b gives the method of calculating the standard deviation from this distribution. For these data (family incomes) $s^2 = 276,290$ and therefore $s = 526$. The value of the mean is 1629.1, and the percentage standard deviation is therefore $100 \times 526/1629.1 = 32.3$. The number required in

the sample to give a standard error of 5 per cent. in the estimate of the mean income per family is therefore

$$n = \frac{32 \cdot 3^2}{5^2} = 42$$

For a standard error of 1 per cent. the number required would be 1050.

A similar calculation for the wheat acreages of the random sample of Hertfordshire farms (Table 7.2), already described in Section 3.7, is given in Example 7.2.b. In this case $s = \sqrt{1351 \cdot 4} = 36 \cdot 8$. The mean wheat acreage per farm is 18.6, and the percentage standard deviation is therefore $100 \times 36 \cdot 8 / 18 \cdot 6 = 198$. The percentage standard error is here very large because there are a large number of farms growing little or no wheat. In order to determine the total wheat acreage of an area with a 5 per cent. standard error from a random sample of farms, therefore, we shall require $198^2 / 5^2 = 1570$ farms.

From the above formulæ we see that the standard errors of estimates derived from random samples of different sizes taken from the same population are inversely proportional to the square roots of the numbers in the samples. Conversely, to reduce the standard errors of the results in a given ratio we require to increase the size of the sample by the square of the ratio. Thus, in order to halve the standard errors of the results we must multiply the size of the sample by 4.

4.32 Some general rules on size of sample

From the above discussion it will be seen that the calculation of the size of sample required to attain a given accuracy is a relatively simple matter when a random sample is taken. With the more involved types of sampling the calculations are more complicated, and more must be known of the material that is being sampled.

Calculation of the accuracy which would be attained by a random sample is, however, often a useful preliminary guide to the size of sample likely to be required in the more involved types of sampling. If only one type of sampling unit is under consideration, the reduction in numbers of units required with the more complicated types of sampling is determined by the fraction of the total variability which is removed by the imposition of restrictions such as stratification or by the use of supplementary information. It is frequently possible to form a rough idea of the likely reduction from a general knowledge of the characteristics of the material. Thus in a survey designed to determine crop acreages, using farms as sampling units, it is to be expected that stratification by size of farm and the use of a variable sampling fraction in conjunction with such stratification will each give considerable increase in accuracy over a random sample. This is confirmed by the results already given in Section 3.7.

When sampling units of different types or of alternative sizes are under consideration, the situation is more complicated, as is shown by the results already presented in Section 3.11. This is true also of multi-stage sampling.

The following general rules may be of value. Rules 1 to 5 are applicable to the case in which only one type of sampling unit is under consideration, rules 6 and 7 to the case in which more than one type of sampling unit is being considered.

- (1) The use of stratification, a variable sampling fraction or supplementary information may in general be expected to increase the accuracy. Consequently the calculation of the number of units required in the case of a random sample gives an upper limit to the number of units required in any reasonable form of sampling using the same sampling units.
- (2) Stratification will only increase the accuracy substantially if there are marked differences between the different strata. The increases are usually larger for quantitative characters than for qualitative characters, *i.e.* attributes (Table 3.7.b).
- (3) A variable sampling fraction can greatly increase the accuracy when the units vary greatly in size, or more generally in variability from stratum to stratum. Fractions which increase the accuracy for quantitative characters may reduce it for qualitative characters (Table 3.7.b).
- (4) The use of supplementary information can greatly increase the accuracy in appropriate cases, and often serves as an alternative to stratification (Table 3.7.b).
- (5) Since there must be at least one unit per stratum, more detailed stratification is possible with larger samples. In such circumstances the increase in accuracy with increasing size of sample will be more rapid than is indicated by the square-root law. Conversely, for samples of a given accuracy the advantage of stratification may be reduced by the fact that reduction in the size of the sample necessitates an increase in the size of the strata (Section 8.15).
- (6) If sampling units of type A consist of aggregates of sampling units of type B (*e.g.* households and individuals), the use of sampling units of type A in place of units of type B will usually result in lower accuracy for a given amount of material in the sample (Table 3.11.b and Example 7.8.b).
- (7) If multi-stage sampling is used, more final-stage units will be required than will be the case with single-stage sampling of the final-stage units (Tables 3.7.b and 3.11.b).

All the above rules are indicative only. The quantitative gains in accuracy or reduction in number of units required in any particular case must be evaluated by the methods described in Chapter 8. The final decision as to the type of sampling to be adopted necessarily depends on the relative accuracy of the various methods and their relative costs.

It is advisable at the planning stage to consider as far as possible the form in which the results require to be presented. In more complicated surveys,

particularly of the exploratory and research type, the results themselves will in part suggest the form in which they require to be presented, but in simple types of survey the form of presentation can often be laid down in considerable detail. This is a help in verifying that the sample is sufficiently large to cover the required domains of study adequately.

4.33 Pilot and exploratory surveys

From what has been said in the previous sections of this chapter it will be apparent that there are many points on which decisions can only properly be reached after preliminary investigations in the form of a *pilot survey* have been carried out. On material of which nothing is initially known, *e.g.* in surveys of undeveloped territory, a preliminary *exploratory survey* may be required before any proper pilot survey can be undertaken. In addition to providing general information, such an exploratory survey may be used to construct a first-stage frame.

A pilot survey has two main objects: firstly, the provision of information on the various components of variability to which the material is subject, and secondly the development of field procedure, the testing of questionnaires and the training of investigators. A pilot survey may also provide data for the estimation of the various components of cost of the different operations involved in the survey, *e.g.* interview time, time of travel, etc. Knowledge of such costs is required not only as a basis for general estimates of cost, but also in order to determine what type and intensity of sampling will be most efficient.

A further function of pilot surveys is to determine the most effective type and size of sampling unit. In a crop-cutting survey involving the harvesting of small areas, for instance, we may require to determine the best size and shape for these areas. In order to investigate the variability of different types and sizes of unit it is necessary to be able to form aggregates which represent the largest units which are of interest. Thus, if areas ranging from, say, $\frac{1}{2}$ ft. $\times$ 1 ft. to 3 ft. $\times$ 3 ft. are under consideration, it is necessary, or at least preferable, to harvest randomly distributed areas of 3 ft. $\times$ 3 ft. in sections of $\frac{1}{2}$ ft. $\times$ 1 ft. (Section 8.14).

Pilot surveys will not normally be required for material on which there is considerable previous survey experience, since every survey provides information on the variability of the material surveyed, and this can often be used in the planning of further surveys. Thus the 1942 Census of Woodlands (Section 4.25) was planned on the basis of experience gained in the 1938-9 Census of Woodlands. Even in such cases, however, new questionnaires and new methods of observation and measurement should be tested on a more or less random sample of the material before being put into operation.

The testing of the field procedure by means of a pilot survey is discussed in Chapter 5, and its planning needs no special comment. The planning of a pilot survey to provide relevant information on the various components of variation is rather more difficult. The finer points will be fully apparent after

the estimation of efficiency has been discussed (Chapter 8), but a few fundamental points may be made here.

At first sight it might be thought that a fully random sample would be a satisfactory form of sample for a pilot survey. This, however, is not necessarily the case. As a simple example we may consider the survey of material in which the use of a stratified sample is likely to be appropriate. In this case we shall require to determine the components of variation within strata. This can only be done from a fully random sample if the sample is sufficiently large relative to the number of strata for the majority of strata to contain at least two units.

This difficulty can be overcome by adopting some form of multi-stage sampling, so that the whole of the pilot sample is concentrated in a few of the strata. The primary stage of this multi-stage process need not necessarily be very rigorous. Thus in a survey covering a human population concentrated in villages, it may be considerably more convenient to use certain villages for the pilot survey rather than others. Provided that sufficient is known of these villages to indicate that they are fairly typical there is no serious objection to their use. Similarly in area sampling it may be sufficient to confine the pilot survey to districts conveniently situated with regard to the main and regional headquarters, provided there is some assurance that the different types of district are properly represented.

Within the towns or areas selected for the pilot survey a fairly intensive survey can be made. If necessary a further stage may be introduced into the sampling process. Thus the survey may be confined to a selection of blocks in a city, instead of the whole city being sparsely covered. By this means, it is possible to obtain data which cover selected areas with a density which is of the same order as that which will be adopted in the final survey. If this is done the various possible types and sizes of strata can be effectively investigated.

The concentration of the pilot sample into selected areas should not be pushed to extremes. It is better to have adequate cover of a representative sample of the data than highly detailed cover of small and possibly non-representative sections of it. Detailed cover will in any case not be required if the material is such that very small strata are known to be impracticable or of no value.

When the possibilities of multi-stage sampling have to be investigated, the problem of designing a pilot survey is more difficult. The component of variability that governs the sampling error for the whole population is the variability of first-stage units (Section 7.17). Only if an adequate number of first-stage units are represented in the pilot sample will it be possible to make any reliable estimate of their variability. Moreover the first-stage units must be selected at random from within the first-stage strata that will be finally adopted.

For this reason a much more extensive pilot survey is required for a multi-stage survey if any reliable preliminary estimate of the sampling error is to

be obtained. In surveys in which comparatively few first-stage units are used, such as localized surveys on human populations (Section 4.18), the prior determination of the expected accuracy from pilot survey data is likely to prove to be impracticable. Reliance will then have to be placed on previous experience or on estimates of error from previously available data. This, however, is not quite so serious as it appears at first sight, since multi-stage samples with relatively few first-stage units are in general used most frequently in surveys of the research and investigational type, or in surveys which are repeated at intervals. In such cases the first few surveys can be used to provide data on the various components of variation, and the design can be modified if necessary in the light of this information.

Even if the first-stage sampling error cannot be determined by means of a pilot survey, such a survey can be made to furnish reliable information on the errors to be expected at the second and subsequent stages. This will enable the survey to be planned so that the necessary accuracy is obtained on comparisons between first-stage units, which frequently form important domains of study in surveys of this type.

Elaborate pilot surveys are not likely to be worth while in small-scale surveys. It is usually better to proceed with the actual survey work, even if the design adopted is not as efficient as would be possible if a full-scale pilot survey were first undertaken. If a series of small-scale surveys of similar type have to be undertaken the earlier surveys will themselves act as pilot surveys for the later surveys, the design of which can be modified in the light of the experience gained.

PROBLEMS ARISING IN THE EXECUTION AND ANALYSIS OF A SURVEY

5.1 Types of problem

The problems arising in the execution of sample censuses and surveys are for the most part similar to those encountered in complete censuses. We shall therefore not discuss these problems in detail, but merely draw attention to some of the points which are of particular importance when sampling is used.

The various phases of the work subsequent to the planning stage may be broadly classified as follows:—

- (1) Setting up of the general administrative organization.
- (2) Design of forms.
- (3) Selection, training and supervision of the field investigators.
- (4) Control of the accuracy of the field work.
- (5) Arrangements for follow-up in the case of non-response.
- (6) Abstraction and coding of the information.
- (7) Statistical analysis.
- (8) Reporting.

In certain types of survey no action may be required under some of these heads. In a survey conducted by postal questionnaire, for example, there will be no field investigators unless they are required to deal with cases of non-response.

5.2 Administrative organization

The administrative organization required will depend very much on the nature and scale of the census or survey, and on the area to be covered.

The main field task for which an extensive administrative organization is required is the supervision of the investigators, or the carrying out of follow-up enquiries in cases in which there is no staff of investigators to undertake this work. The main administrative task at headquarters is the supervision of the computing and clerical staff engaged in the abstraction and analysis of the completed forms.

Every opportunity should be taken to utilize existing administrative and office organizations. When the survey covers a large area, supervision from a central office is likely to be difficult and in such cases it is best to establish regional offices. Very frequently some existing organization can be used for this purpose.

It is not necessary for the computing and clerical staff at headquarters to be administered by the same organization as the field staff. In many cases it is convenient to use some existing statistical organization to carry out the analysis, and to utilize some administrative organization with regional offices

to supervise the field work. Often an organization can be employed which already has contact with the respondents, or which has on its staff individuals who are suitably qualified to act as field investigators.

5.3 Design of forms

Most careful attention should be given to the detailed design of the various forms that will be used in the course of the census or survey, especially the forms on which the observations and answers to questions are recorded. This applies also to the instructions and explanatory notes which accompany the forms.

The content of the forms for the recording of the information is determined by the information that is required, and has already been discussed. They may be forms designed for completion by the recipients with little or no assistance, questionnaires which form the basis of interviews, or forms on which observations and measurements taken in the field are recorded by the field investigators.

Each type of form presents its own difficulties of design. The simplest is that on which observations and physical measurements are recorded by the field investigators themselves. In this case, the chief points to observe are that the form is convenient to use, and that the results are set out in such a manner that they are convenient to abstract. Figures which have to be summed by the field investigators, for example, should be arranged vertically and not horizontally, as the investigators will not be using calculating machines.

In surveys which involve observations and physical measurements it will almost always be necessary to supply field investigators with a separate set of instructions. Consequently there is no need for the form to carry its own full explanation, though it should of course be made as self-explanatory as possible. Experience has shown that instructions to field investigators should be very detailed, and should cover all possible points of uncertainty or ambiguity. Provision should also be made for revision and amendment as need arises, since it is extremely difficult to draw up a set of instructions which are completely unambiguous and deal with all possible contingencies.

In forms of the census type, designed for completion by the recipients without assistance, very careful attention must be paid to the exact wording both of the questions and explanatory notes, so that there is no doubt in the mind of the recipient as to what is required. Detailed and lengthy explanations should be avoided as far as possible. Such explanations as have to be given should if possible appear in conjunction with the question to which they refer. The common practice of giving detailed explanatory notes on the back of a form is not very satisfactory, since it frequently results in the respondent filling in the whole or portions of the form without consulting these notes. Forms of this type should, if possible, carry a brief explanation of the reasons for the census. Even if this has been given in the press and elsewhere it is unlikely that all recipients will in fact have seen it.

In forms of the questionnaire type designed for completion by field investigators the investigators must be instructed whether the questions are to be put in the exact form given, or whether they can be asked in a general form. As already stated, in most cases the general form is more suitable, but in questions on opinions, where different forms of wording may be expected to affect the answer, it may be necessary to adhere to an exact form.

With the general form of question explanatory notes are often required in order to make clear to the investigators exactly what information is required. Such explanatory notes can either appear on the questionnaire itself or be given in a separate set of instructions. The latter course results in a much more compact form of questionnaire and is suitable when full-time investigators are used. The former course is more likely to ensure that all investigators are in fact aware of what is really required and is best when the investigators are carrying out the survey in the course of other duties. In a lengthy questionnaire this will necessitate the questionnaire being in the form of a booklet. Such questionnaires are more bulky and costly, and frequently entail more work in the coding of the results, but are nevertheless frequently preferable in these circumstances.

Forms may be either printed or duplicated. Printing is much to be preferred as it results in much neater, clearer, and more compact forms. The ordinary type of duplicating paper is also not very suitable for writing on, particularly in ink.

Small forms may be printed on cards instead of paper. Cards are often more convenient for field use, and in small surveys of which the results are analysed by hand the use of cards may save transcription before analysis. Alternatively Cope-Chat cards may be used (Section 5.10).

Forms printed on paper may be made up in the form of blocks with cardboard backs. This facilitates writing in the field. Alternatively they can be clipped on to a wooden board. If duplicate copies of the completed forms are required, provision should be made for carbon copies to be taken at the time the forms are filled in.

Forms larger than foolscap should be avoided if possible. They are troublesome both to handle and to store. Forms of more than one sheet should also be avoided. It is usually better to use both sides of a sheet or card than to use two sheets.

Forms should always be subjected to a preliminary trial in the field. Only in this way will minor faults be discovered. In the case of questionnaires this test is best arranged in two parts:

- (a) a trial by investigators who are fully experienced in questionnaire work, and who are conversant with the problems under investigation;
- (b) a trial by investigators of the type that are to be employed in the survey.

The first trial will serve to determine whether the questionnaire is in the form most suitable for eliciting the required information from the respondents,

and the second trial will provide information on whether the questions and associated instructions are understood by and within the capability of the investigators.

5.4 Special tests of questionnaires and investigators

In certain cases it may be worth making rigorous tests of different forms of the same question to see whether there are any material differences in the answers received. Since the question cannot be put in both forms to the same respondent, this must be done by the use of interpenetrating samples, using the same investigator or investigators for both forms. In order to eliminate the effect of any progressive change in the investigators or the respondents the tests of the two forms should proceed simultaneously. In the same way the difference between two or more investigators using the same form of question can be tested.

More elaborate and precise tests of differences resulting from different forms of the same question and different investigators can be carried out by using the methods developed in the design of experiments. Thus if two forms P and Q of a question and three investigators A , B and C require to be tested, groups or *blocks* of six respondents may be used. The blocks should be chosen in such a manner that the respondents within each block are as alike as possible, using any available prior information. The six question-investigator combinations PA , QA , PB , QB , PC , QC are then assigned at random to the respondents of each block. This design is technically known as a 2×3 *factorial design in randomized blocks*. By this device differences between forms of question and between investigators are simultaneously tested. Information is also obtained on what are known as the *interactions* between forms of question and investigators, *i.e.* on whether the differences between the forms of question are different for the different investigators, and vice versa. The grouping of respondents into blocks ensures that errors due to differences between respondents are eliminated as far as possible; the randomization enables the standard errors of the comparisons to be calculated by the methods appropriate to the analysis of replicated experiments (see for example Fisher's *Design of Experiments* or Snedecor's *Statistical Methods*).

Investigations of this kind can be carried out in the course of an actual survey, but they are normally better undertaken as a special investigation or as part of the pilot survey, since information on different forms of question will be required at the planning stage, and it is usually inadvisable to complicate the field procedure of a large survey. Routine tests of differences between different investigators may, however, be incorporated without undue complication in the actual survey by means of interpenetrating samples.

5.5 Selection, training and supervision of field investigators

Field investigators may be specially appointed, they may be members of existing staffs appointed for other work but over whom authority can be

exercised, or they may be individuals asked to undertake the work on a voluntary basis or for a small honorarium.

The problem of selection arises primarily in the case of investigators appointed specially for the work. In order to secure a suitable type of person preliminary tests should if possible be made of all applicants, and the early work of newly appointed investigators should be carefully watched and supervised. In large-scale censuses and surveys, proper training courses should be arranged. If a pilot survey is undertaken this provides a valuable opportunity for training, and every attempt should be made to build up the team of investigators at this stage rather than later, even if this involves a certain amount of additional expense.

It is of the greatest importance that investigators, once they have been trained and are found suitable, remain in the job. Every effort must therefore be made to see that the pay is adequate, and that the work is made as attractive as possible. In the case of the interview type of survey, investigators are sometimes paid on piece rates at so much a completed questionnaire. This is in general unsatisfactory, since it tends to lead to skimmed work and to irregularities such as substitution of one respondent for another.

It should not be forgotten that field work of the interview type is very arduous and is found by almost all investigators to involve considerable mental strain. Hours of work are also likely to be irregular, since if excessive non-response is to be avoided some evening interviews are almost inevitable. Investigators should therefore not be expected to work excessively long hours, and should if possible be given a rest on other work from time to time. It is often advantageous to bring full-time investigators to headquarters at intervals and use them for office work such as abstraction and analysis of the results. This not only serves to provide a break from field work, but also enables them to gain a much better insight into the purposes of their work.

Whatever the conditions of work and form of payment, there must be adequate field supervision. The supervisors should themselves undertake field work from time to time, so that they are in a position to appreciate the difficulties of the work, and should also contact the workers while they are actually in the field. Provision should be made for personal contacts not only between supervisors and the field investigators, but also between supervisors and the headquarters staff. In long-term surveys it is also often advantageous to arrange conferences of the investigators from time to time at which difficulties can be discussed and the whole progress of the survey reviewed.

5.6 Control of the accuracy of the field work

The best assurance that the field work shall be accurate is that the investigators are thoroughly trained in their work, and are capable, conscientious, and keen. Nevertheless it is important even with the best investigators to keep a close watch on the progress of the work.

In certain cases, particularly in surveys involving observations and physical

measurements, it is possible to arrange a system of field checks by the supervisors. These should preferably be carried out on a random sub-sample of units, and should in any case be conducted in such a manner that the investigators cannot know which parts of their work will be checked. Checks of this type will not usually be possible in the interview type of survey, as it is clearly impracticable to ask for the same information twice from the same individual.

A preliminary examination of the completed forms must be made as soon as possible after they are completed. In this way defective work, in so far as it reveals itself in the forms themselves, is brought to light immediately, and remedial action can be taken. If the census or survey is such that a large volume of work is turned in by each field investigator, and it is not considered necessary to give individual scrutiny to all the returns, a proper sample of the work of each investigator should be scrutinized as a routine matter.

The investigators should themselves be instructed to carry out any simple numerical calculations that are required on the forms, and also to look through the forms before sending them in to see if they are in satisfactory order. On the other hand extensive revision of the forms should not be permitted. The preparation of fair copies by the investigators is in general undesirable, since it leads to copying errors and also makes any judgment on the quality of the work more difficult. If fair copies are permitted the originals should be returned together with the fair copies, and a certain percentage at least should be checked for copying errors and other changes.

If the questionnaire is such that the investigator has to furnish or amplify the answers to some of the questions from notes taken at the interview this should be done immediately after the interview rather than at the end of the day, even if this course is somewhat inconvenient. In general, however, it is best for the information to be written down in its final form at the interview, any supplementary observations by the investigator being given under a separate heading.

If comparisons between the different investigators by means of interpenetrating samples have been arranged, the comparative results must be made available as quickly as possible, in order that effective action may be taken if discrepancies are discovered. On the other hand the use of interpenetrating samples should not be made an excuse for the relaxation of other forms of control. Interpenetrating samples are not likely to reveal minor defects in an individual investigator, and they will certainly not reveal faults which are common to all investigators. They should therefore be regarded as a check against major defects in individual investigators rather than as a complete control of all investigators.

5.7 Arrangements for follow-up in the case of non-response

The follow-up arrangements will naturally vary very greatly according to the type of census or survey.

In the case of a postal questionnaire they will often involve an entirely different organization from that which is employed to carry out the survey

itself. Since postal follow-ups from headquarters are of limited utility, some form of local organization which can deal with non-respondents by telephone and personal visit is required.

In surveys using field investigators careful instructions must be issued in order to be sure the follow-up arrangements are properly carried out. Specific warnings should be given against such practices as substitution of neighbouring households when there is no response. If the follow-up is to be made on a sub-sample only of the non-respondents exact instructions for taking this sub-sample must be given, so that it can be obtained as soon as the non-respondents are known. In general some very simple sampling method, such as taking of every q th non-respondent, is adequate. Such a procedure has the advantage that a list for follow-up can be prepared at the time of the original non-responses, if necessary by the field investigator concerned.

5.8 Statistical analysis

Statistical analyses of the results of complete censuses and surveys are mostly based on counts of numbers of units falling in different classes and sub-classes, and on the totals for these classes of recorded quantitative variates. The units to which these counts and totals refer may be either the sampling units or some other natural units. In certain cases totals are also required for ratios or other quantities calculated from the values recorded for the individual units.

From these numbers and totals the means can be calculated for the different classes. Basic summary tables can then be prepared. In these summary tables frequencies based on counts are often expressed in percentages, the bases of the percentages being chosen so as to exhibit the differences in proportions which are of interest. When the summary tables have been prepared, more critical statistical analysis of these tables may be required in order to isolate the effects of the various factors which are believed to influence the results.

The treatment of the results of sample censuses and surveys is similar in most respects to that of complete censuses and surveys. If, however, the sampling fractions are different for the different units, the appropriate weights have to be applied at some stage of the calculations. The utilization of supplementary information will also necessitate adjustment of the basic totals and means. The appropriate formulæ for these operations, which differ according to the type of sampling adopted, are given in Chapter 6.

In sample censuses and surveys we shall normally require estimates of the sampling errors. In addition, investigations of the relative efficiency of different sampling methods may be undertaken in order to improve the efficiency of future surveys on the same or similar material.

The various stages in the computations may therefore be classified as follows :—

- (1) Preliminary computations on the values of the individual returns, such as the calculation of ratios and the introduction of individual weighting

factors. If punched cards are used, some or all of this work may be carried out mechanically, subsequent to the punching of the cards.

(2) Abstraction and coding of the results so that they are in a form suitable for analysis or for transfer to punched cards.

(3) Punching (in cases where punched cards are used).

(4) Counts and totals.

(5) Preparation of the summary tables from these counts and totals, including adjustments for supplementary information and any weighting not already carried out.

(6) Calculation of sampling errors and investigations of efficiency.

(7) Critical analysis.

Apart from the calculation of sampling errors and investigations of efficiency, which are described in Chapters 7 and 8, we do not propose to discuss these operations in detail in this book. In the following sections we will merely give an outline of the special points that arise at the various stages.

5.9 Methods of handling the data

There are four main ways in which the data accumulated in the course of a census or survey may be handled. These are:

(1) An analysis direct from the forms.

(2) Transference of the data to ordinary cards.

(3) The use of cards with holes round the edges (Cope-Chat cards).

(4) The use of Hollerith or Powers-Samas cards (punched cards).

The primary function of any type of card is to enable the data to be sorted into different classes, so that the numbers of units and totals associated with these classes can be obtained without transcription. With plain cards the sorting has to be done entirely by hand, with Cope-Chat cards marginal punching gives some aid to the hand sorting process, while with punched cards the sorting is carried out mechanically, and the counts and totals are also obtained mechanically.

In certain cases the data can be recorded directly on cards which are subsequently used in the analysis. These may be either ordinary cards or Cope-Chat cards. The use of cards in this manner is limited by the fact that the amount of uncoded information that can be conveniently recorded on a card is small, and also by the fact that cards tend to be damaged by use in the field.

It is also possible to record information directly on Hollerith cards, either in a form which enables it to be read by the punch operator as the card is punched, or in a form that enables it to be punched automatically by the process known as *mark sensing*. The occasions on which either of these methods has any real advantage over the punching of cards from ordinary forms are somewhat rare in census and survey work.

If only a single classification is required, the preparation of a summary directly from the forms is likely to be the most economical method of procedure.

If more than one classification is required, the use of forms may still be reasonably economical, particularly for small surveys, but the possibilities of using forms in this manner are limited by the fact that paper forms are not easily sorted or counted, and will not stand a great deal of handling.* The direct use of forms is also sometimes of value when a rapid preliminary summary of the salient features of a survey is required. In a large survey such summaries can usually be based on a small sub-sample of the forms.

If the data are transferred to cards, some form of compression and coding is usually necessary. This enables the information to be recorded in compact form on the card, and also facilitates the subsequent counts and summation. If Cope-Chat cards are used, all information recorded in punched form must be coded, and with punched cards the whole of the information has to be coded in numerical (or exceptionally alphabetical) form.

The fact that punched cards have all their information coded in numerical form has the disadvantage that the detailed information relating to separate units cannot be easily studied by means of the cards themselves. It is also difficult to record written remarks on the cards. This tends to make the analysis more mechanical. Punched cards are therefore unsuitable for analyses which require detailed examination of the whole complex of information relating to individual units. Even in surveys which are so large that analysis by means of punched cards is essential it is often advisable to arrange that the original forms are kept available, so that in any detailed investigational work the forms corresponding to selected cards can be extracted and examined when required.

5.10 Cope-Chat cards

Cope-Chat cards are cards which have a row of holes along each edge. A group of these holes can be assigned to each particular classification, *e.g.* the answers to a specific question, each hole being taken to represent one class in this classification. The body of the card (front and back) can be used for recording written information.

By means of a punch similar to an ordinary ticket punch, V-shaped notches can be cut out of the card so as to obliterate any desired holes. If the cards are arranged in a pack and a knitting needle is passed through a particular hole, the cards punched in this hole will fall from the pack when the pack is lifted by means of the needle and thoroughly shaken. This enables cards to be sorted into different classes with considerably greater speed than would be the case if the information were merely recorded on plain cards, and the sorting had to be carried out by examination of each card. The Cope-Chat method of sorting is not fully reliable, since cards do not always fall out of the pack when it is shaken, but mis-sorts can be detected by visual inspection of the edges of the retained cards. If all classes of a given classification are coded in some mutually exclusive system a positive check will be available.

* In making counts or calculating totals from forms it is usually best to sort the forms into the necessary classes.

The amount of information that can be recorded on the edge of the cards is limited, since the number of holes is limited by the size of the card. In the Survey of Fertilizer Practice, for example, 5 in. $\times$ 8 in. cards are used with a hole spacing of just over 4 to the inch, giving 105 holes in all.*

The punching of Cope-Chat cards is somewhat laborious, and if a mistake is made a new card has to be prepared. In this case the written information will have to be transferred to the new card. For this reason it is usual to mark the holes which have to be punched, and check the markings before actually punching. If a considerable amount of punching is to be done, a form of gang punch can be used which will punch a particular hole from a number of cards at one operation. In this case the cards can be sorted into the appropriate classes and the sorting checked before punching. A key-operated punch is also available.

When the cards have been sorted they require to be counted by hand. If totals of numerical information are required, the summations must be performed on an ordinary adding or calculating machine, unless the numerical information has itself been coded. The counting of cards is a tedious operation, and is made more so by the punching round the edges. For some purposes it may be feasible to replace exact counts either by weighing or by measuring the aggregate thickness under a definite pressure. Neither method is very accurate, however. In a humid climate, for example, the weights tend to vary considerably owing to changes in moisture content.

The use of Cope-Chat cards enables isolated cards having given characteristics to be much more readily extracted than is the case when plain cards are used. Cope-Chat cards are therefore of value for surveys in which units of particular types require to be identified subsequently. In this respect they have certain advantages over punched cards, since no elaborate sorting mechanism is required and the information concerning the selected units is presented in written form.

Cope-Chat cards also have the minor advantage that the proportions falling in different classes can be roughly observed by sorting the cards and then examining the distribution of the notches.

The coding of numerical information on Cope-Chat cards can be carried out in a number of ways. If approximate values only are required the data may be grouped into size-groups. If exact values are required the simplest method is to allocate ten holes to each digit of the number, but this can only be done if very little numerical information has to be coded, owing to the limited capacity of the card. An alternative is to use some form of two-hole code to represent each digit. The most compact is that based on four holes, which are taken to denote the digits 1, 2, 4, 7. To code other digits the two digits whose sum gives the required digit are punched. Thus the punching of 1 and 2 represents digit 3. This system is not self-checking on sorts, since

* In both the Cope-Chat and punched card systems one corner is cut across diagonally on all cards so as to provide a check that all cards are right way round in the pack.

two, one or no holes may be punched. If a fifth hole carrying the value 0 is added, every digit can be denoted by a pair of holes, with the convention that 4, 7 denotes 0. An alternative with five holes, which is simpler, but not self-checking on sorts, is to use the holes to denote the digits 1, 2, 3, 4, 5, digits over 5 being indicated by double punching.

5.11 Punched cards

Two different systems are available, known as the Hollerith and Powers-Samas. Both systems employ cards in which each column has 12 positions, in any one of which a hole may be punched. When required, two or more holes may be punched in different positions in the same column. In both systems alphabetical information can be dealt with by means of a two-hole code.

Hollerith installations employ 80- or 38-column cards. Powers installations employ 65-, 36- or 21-column cards. A given installation will only handle cards of one size. By using each column for two items of information, with a special form of multiple punching, the Powers 65-column card can be extended to give the equivalent of a 130-column card.

The actual punching of the cards is normally done by a hand-operated key punch. Verification, which checks within certain limitations that the original punching is correct, is normally performed by means of a hand-operated verifier similar in construction to a punch. More elaborate punches of various kinds are also available.

The main difference between the Hollerith and Powers systems is that in the Hollerith system the cards are read electrically, whereas in the Powers system they are read mechanically. This results in a greater flexibility in the Hollerith system, since the machines can be set up for any required operation by means of electric connections through one or more plug-boards. If the analysis is confined to sorting and counting, the two systems, apart from card capacity, have almost identical performance. For the more elaborate types of analysis, Hollerith equipment is more suitable than Powers equipment, particularly in surveys of moderate size where many different types of machine operation, which often cannot be planned in advance, are required on relatively small batches of cards. We shall here confine ourselves to a description of the Hollerith machines, but it should be emphasized that if Hollerith equipment is not available it may be preferable to utilize existing Powers equipment rather than send the work elsewhere or use methods not involving punched cards.

The principal Hollerith machines are :—

- (1) The sorter,
- (2) The sorter-counter,
- (3) The tabulator,
- (4) The reproducing summary punch,
- (5) The multiplying punch,
- (6) The collator.

The descriptions which follow are not intended to give a complete account of these machines and their various modifications, but only an indication of the way in which they work and the simpler types of operation that can be undertaken with them. An expert should always be consulted when planning any extensive punched-card work. Initial consultations should take place before the coding of the material is undertaken.

5.12 The sorter and sorter-counter

The sorter can be set to operate on any one column of the card. When the cards are passed through the machine they are separated into 12 boxes corresponding to the 12 positions of the holes punched in this column, with an additional box for cards with no hole in the column. If, therefore, a code representing some classification of the material into anything up to 12 classes is punched in the column, the cards corresponding to the different classes will be sorted into the different boxes. A classification with more than 12 and up to 144 classes can be coded on two columns, and by sorting successively on each of the two columns separation into all the classes can be effected. In the same way, if a group of columns or *field* is used to denote a number, the cards can be arranged in numerical order by sorting first on the units, then on the tens, and so on. Equally, if two columns represent two different classifications the cards can be sorted into the various cells of the two-way classification so formed. If two holes are punched in the same column the card is sorted to the higher digit, unless sorting on this digit is suppressed.

The sorter is normally used for arranging the cards of the pack into groups or into a given order prior to their passage through the tabulator. When this is done the whole of the cards are kept in one pack, *i.e.* at the end of each sort the cards are collected from the separate boxes and the sub-packs are placed together in numerical order.

If only counts are required it is possible to obtain these directly on a sorter with a counter device which registers the numbers of holes occupying the various positions in the given column. A machine with this device is called a *sorter-counter*. Sorting can be suspended during counting if desired.

The ordinary sorter-counter counts on a single column only and does not print the results. For large-scale census work more elaborate types of sorter-counter are available which will count simultaneously on a number of columns, printing the results obtained in these counts.

5.13 The tabulator

The tabulator is a much more elaborate machine than the sorter. Its primary function is to add numbers punched in a given field from a group of cards. To effect this the numbers are read successively as the cards pass through the machine, being added on one of a set of counters which form part of the machine. The machine has a printing device which will print the totals accumulated in the counters, and will also, if desired, print numbers read from

the cards. The operation of obtaining and printing the totals is called *tabulation*, and that of printing numbers read from the cards is called *listing*.

Most tabulators have a number of counters and print banks ; the totals of several fields can therefore be accumulated simultaneously. If the numbers in the fields concerned are sufficiently small, two or more fields can be accumulated in different parts of the same counter, thereby further increasing the capacity of the machine.

In order to enable the totals of the groups of cards in different classes to be obtained successively without stopping the machine and without having to feed in the groups of cards separately, a device known as the *control* is incorporated in the machine. This device is such that when wired to control on a given column, the machine will break control if the card following the one that is being added carries a different designation on the control column. This break of control stops the adding process and gives the machine certain instructions as to printing and clearing, *e.g.* it can be wired so that the total already obtained is printed and cleared before passing on to the next group of cards. Thus, if a pack of cards sorted into groups corresponding to the code on a single column is passed through the tabulator with the control wired to that column, the machine will break control at the end of each group and the group totals of any desired field can thereby be obtained.

The control can be arranged to operate on a number of columns, and different stages of the control can be associated with the different columns. Different instructions can be given to the clearing and printing mechanisms according to which stage of the control is operating. Thus, for example, it is possible to obtain totals of main and sub-groups simultaneously by feeding the numbers from the given field into two different counters, one of which is cleared at the end of each sub-group and the other at the end of each main group.

Counts can be carried out on the tabulator, either in conjunction with a tabulation or independently, by what is known as the card count. This feeds 1 into any desired counter at the passage of each card. The control and printing mechanisms operate as before.

The more elaborate forms of tabulator have a number of auxiliary devices which considerably increase their potentialities and flexibility. The two most important in the British machines are the rolling feature and *distributors*. In the rolling total tabulator, numbers can be transferred or *rolled* from one counter to another, either positively or negatively, according to instructions issued by the control mechanism. Distributors enable numbers read from a field to be directed to different counters, and also enable numbers taken from one counter to be directed to different counters in rolling, or to different print banks. When used in the first manner the distributors operate on instructions read from some other column of the card.

A single distributor, for example, enables positive and negative numbers in the same field to be distributed into two counters according to their sign (punched in code in another column). By rolling the total of the negative

counter negatively into the positive counter at the end of the group the correct total (or its complement if negative) is obtained. A further device replaces the complement by the negative total on printing.

By using four distributors it is possible to make counts of classes represented by a code in a single column without sorting on that column. In this case the card count is fed through the distributors which are controlled by the punching of the column in question. The wiring is so arranged that the card count is directed to a different counter or part of a counter for each of the 12 possible code punchings.

The use of a tabulator in place of a sorter-counter for counting, either with or without distributors, has the advantage that the results are obtained in printed form. It also enables the whole pack of cards to be kept together, since the different main classes are automatically separated by means of the control.

It should be noted, however, that when there is multiple punching in the column being counted, a tabulator performs a somewhat different operation from the sorter-counter. A sorter-counter counts all holes punched in a column, e.g. if 4 and 8 are punched, both the 4 and the 8 will be counted. With a tabulator only one of the holes is counted, and consequently additional operations are required in order to obtain a full count.

Rolling can be used to carry out simple multiplications. In the National Farm Survey analysis, for example, a variable sampling fraction with values $1/20$, $1/10$, $1/4$, $1/2$, $1/1$, was used for the different size-groups. Multiplication of the size-group totals by their appropriate raising factors was effected as follows:—

- (a) For multiplication by 2 the total was rolled into itself.
- (b) For multiplication by 4 operation (a) was repeated.
- (c) For multiplication by 10 the total was rolled through a distributor so as to give transposition by one place.
- (d) For multiplication by 20 operations (a) and (c) were combined.

Sums of squares and products, which are required for the estimation of sampling errors, regression coefficients, etc., can be obtained on a tabulator by what is known as *progressive digitting*. Suppose the sum of the products of two sets of numbers A and B is required, and that all the A 's are single digit numbers. The cards are sorted on A and then fed through the tabulator, 9's first, the B 's being summed and the progressive total printed (without clearing) at each change of digit of A . The sum of these progressive totals (excluding the final or 0 total) gives the sum of the products of A and B . On a rolling total tabulator this summation can be carried out on the tabulator, provided steps are taken by the insertion of extra cards to see that every digit is represented. If the A 's contain more than one digit each digit is treated separately, with multiplication by 10, etc., before the results are combined. The whole of this operation can be effected in full on the larger rolling total tabulators. Subject to limitations of capacity, sums of products of A with

itself and several other numbers *B*, *C*, *D*, etc., can be carried out simultaneously without additional sorting or tabulation.

The American tabulators do not have the rolling feature, but are capable of carrying out direct addition or subtraction according to card designation. In contrast to the rolling total tabulator, the counting wheels may be grouped in counters entirely at will, which enables the counter capacity to be used more efficiently. These tabulators, however, have no distributors, which detracts somewhat from their usefulness in the analysis of survey data.

5.14 The reproducing summary punch

The reproducing summary punch has two main functions. One is reproducing information from a pack of cards on to the corresponding cards of another pack. The second is what is called *gang-punching*, that is, the punching of information, read from the first or *master* card of the group, on the whole of a group of cards. The punch can also be used in association with a tabulator to punch on to new cards the results obtained in the course of a tabulation.

In reproduction the information in any set of columns can be transferred to the same or any other set of columns. In gang-punching the information will be punched in the position in which it appears on the first card of the group.

The reproduction of information from one pack on to another is of value in survey work in a number of ways. In addition to the obvious function of making a new pack of cards when an old one has become worn, it can be used to bring together on to a single card items of information referring to the same unit and recorded on two or more separate cards, so that the association between these items can be analysed. It also provides a satisfactory means of entering new information on to cards that have already been punched. Instead of punching the new information on the old cards directly, this information is punched on to a new pack, together with the code numbers of the units, and the information from the new pack is then transferred to vacant columns on the old pack or vice versa by means of the reproducing punch.

When transferring information from one pack to another pack which itself already carries information, the two packs must of course be sorted into the same order. The machine, however, checks that there is correct correspondence between each pair of cards, and also checks that all transferred information is correctly punched.

Gang-punching can be used to save hand punching when a batch of cards which are to be punched all carry the same code in a number of columns. It can also be used to transfer information from the main card on to secondary cards or *trailers* referring to the same unit. In this case the main cards act as master cards. In gang-punching with interspersed master cards, the master cards must carry an *X* in one of the columns in which none of the remainder of the cards carry an *X*. If such an *X* is not already punched it can be gang-punched.

A further use in survey work is the calculation of percentages, index numbers, etc. A simple example will illustrate the procedure. Suppose it is required to express the numbers B as a percentage of the numbers A , both sets of numbers being punched on the cards. By suitable sorts we can assemble the cards in batches such that the percentage value for all the cards of each batch is the same. Cards for which A has the value 45, for example, and a value of B between 0 and 2 will have a percentage value of 0 (to the nearest 10 per cent.), those with B between 3 and 6 will have a percentage value of 10, etc. Master cards are therefore made out carrying the values

A	B	Percentage
45	0	0
45	3	10
45	7	20

etc., with similar sets for other values of A , and these are added to the pack, which is then sorted into numerical order of the A 's, and of the B 's within A 's. All the A 's having a value of 45, for example, will now be together, those with B between 0 and 2 being preceded by the first of the above master cards, those with B between 3 and 6 by the second, etc. If the whole pack is then passed through the reproducing punch the correct values of the percentages will be gang-punched into the remaining cards from the master cards.

The disadvantage of this procedure is that a large number of master cards are required to cover with any high degree of accuracy fields which are at all extensive. Time and expense is therefore involved in the preparation of the cards, and they also add to the total volume of sorting required. The method is therefore most suitable when ratios and indices of low accuracy are required for large batches of data. In the analysis of the National Farm Survey it was used for the calculation of the percentage of the acreage of individual farms which was arable (12 classes), rent per acre (12 classes), etc., and for the combination of several items of qualitative information into a single index. A description of the procedure, and a method of preparing master cards by use of the gang punch, is given by Kempthorne (1946, B).

5.15 The multiplying punch

The multiplying punch is designed to read two numbers from a card, calculate the product and punch the result, with suitable rounding-off, in any field of the same card. It can also be set to read the multiplier from interspersed master cards, so that the numbers on the whole of a group of cards are multiplied by the same factor. Cross-footing multiplying punches will also add or subtract two or three numbers read from the same card and punch the result, or add one or two numbers to the product of two other numbers.

The chief use of the multiplying punch in survey work is in the calculation of products of various kinds prior to summation. It can be used for the calculation of ratios if the reciprocals of the divisors are punched on master cards and the cards are then sorted according to their divisors. It is, however,

relatively slow in operation. Consequently when a large amount of material has to be handled and high accuracy is not required in the products or ratios, the use of gang-punching with interspersed master cards is generally to be preferred.

5.16 The collator

The collator will take two packs previously sorted in numerical order on up to 16 columns and will combine them into a single pack, so arranged that all cards of pack *B* which carry a given code-number follow immediately on the cards of pack *A* carrying the same code-number. It will also select matching cards from pack *A* and pack *B*, rejecting cards of pack *A* with code-numbers which do not occur in pack *B* and vice versa. Furthermore it can be used to select cards from one pack which correspond in code designation to the cards contained in a second pack. This last property is sometimes of value in survey work, since it can be used to pick out trailers associated with a given set of main cards. If it is so used, however, the precaution should be taken of matching up the rest of the trailers with the remaining main cards so as to provide a check that no trailers have been erroneously excluded.

5.17 Systems of coding for punched cards

When punched cards are used all non-numerical information will require to be coded in some quasi-numerical form. Each column of the card can represent a classification of up to twelve classes, which are denoted by X, Y, 0, 1, 2, . . . 9.

Numerical quantities do not in general require to be coded, but may require rounding-off in order to economize card space and reduce the counter capacity needed in the subsequent tabulations. Rounding-off, however, is a tiresome operation and reduction of a number by a single digit should only be undertaken if really necessary. Often rounding-off can be avoided by issuing suitable instructions to the field investigators regarding the number of figures required in the results.

In certain cases it may pay to code a numerical quantity by grouping. This is particularly advantageous when the quantity is primarily required as a basis of classification and does not need to be summed. If summation is needed, large values must not be too coarsely grouped, and there must be no "open" group: it is not sufficient for all the high values to be included in an "over —" class. This often necessitates an additional column or over-punching in the X and Y positions. If there is to be summation the grouping must also be chosen so as to avoid the bias which can arise through the frequent occurrence of particular values (see Example 7.2.b), though such bias is relatively unimportant when only comparative results are required.

The construction of a code for complicated items of information, *e.g.* questions with a large number of possible alternative answers, is often a difficult task, since the conflicting aims of recording the information adequately,

simplifying the actual task of coding, and keeping the code compact have to be reconciled. In many cases the most suitable form of coding for a given item of information can only be devised by examination of a sample of the questionnaires or returns.

Compactness of coding is important not only in saving card space, but also in simplifying the tasks of sorting and grouping the cards into classes. With a two-column code, other than the type described in the next paragraph, both sorting and counting are more complicated than in a one-column code. Furthermore, classes which have been separately coded cannot be grouped for purposes of tabulation unless a group code is gang-punched or control is omitted from the code columns, since the control will automatically break at each change of code.

For this reason it often pays to code an item of information in two parts, main and subsidiary. Thus in an agricultural survey of England and Wales the place location will be given by one of 61 counties or part counties. Instead of numbering the counties consecutively the provinces can be numbered, and the counties numbered consecutively within provinces. This will facilitate sorting and tabulation of the material by provinces, and the code will still only require two columns.

Provision should always be made in a coding scheme for recording lack of information on items for which this contingency is likely to arise. Leaving the column in question blank is not satisfactory, as every occupied column should be punched on all cards. For the same reason numbers of under 100 in a three-column field, for example, should commence with 0, not a blank.

If there are a large number of questions to which only two or three alternative answers are possible the columns required on the punched card can be reduced by combining the answers to two or more questions in a single code. Thus two questions, each with the alternative answers yes, no, don't know, can be coded in one column by using the 9 combinations of the answers. Such coding, however, is decidedly more troublesome and liable to error, and is also less convenient for subsequent analysis. In large-scale surveys, therefore, it is often better to keep such questions separate even if this means using an additional card.

In questions in which the answers are not mutually exclusive, multiple answers are usually recorded by multiple punching, *i.e.* punching in a single column the holes corresponding to all the answers given.

Multiple punching can also be used to economize card space in other ways. A two-hole code, for example, is used for alphabetical coding. Unimportant classifications, if they contain sufficiently few classes, can be punched on different parts of the same column. This, however, should only be done if it is reasonably certain that only counts of these classifications will be required, and that they will not have to be used for the control of tabulations.

In cases in which the majority of numbers in a field are below, say, 1000, but there are a few numbers which are between 1000 and 3000, numbers

between 1000 and 1999 can be denoted by overpunching X in the first column of the field and numbers between 2000 and 2999 by overpunching Y. Certain tabulators are fitted with a device (the *29 feature*) which enables numbers so punched to be added directly in the course of the tabulation. If such a device is not available the X's and Y's will have to be separately counted and adjustments made.

Multiple punching should not be used excessively. It slows up the punching, is difficult to verify, and introduces complications into the sorting and tabulations. If the data cannot conveniently be coded in some simple form that will go on a single card with little multiple punching, it is usually better to use an additional card, recombining the data as required by means of a reproducing punch.

A way of avoiding multiple punching when dealing with occasional numbers which exceed the allotted capacity of the field concerned is the use of *trailers*. Thus a number 2571 can be recorded in a three-column field by the use of two trailers, the numbers 999, 999 and 573 being punched. Apart from these numbers only the code number need be hand-punched on the trailers, the remainder of the information being gang-punched from the main card. Trailers must be distinguished from main cards. If this is done by punching 0 and 1 respectively on some column the 1 can be used for counts; this, however, requires an additional column. If X and Y are used in some occupied column an additional distributor will be required for counts. Alternatively the trailers can be removed when counting.

Simple qualitative information can often be *pre-coded* on the questionnaire form. Thus a question to which the only alternative answers are yes, no, don't know, can have these answers printed on the form in conjunction with the numbers 1, 2, 3. The investigator is then instructed to ring the appropriate code number. If it is necessary to make provision for possible non-standard answers partial pre-coding can be used, a line being left for alternative answers which are subsequently coded in the office.

Pre-coding has the advantage that the amount of office work is considerably reduced, since the forms can be sent for punching after scrutiny without further work. It must, however, be confined to questions to which the alternative answers can be printed on the form. It is unsuitable in the case of questions to which complicated and involved answers are likely to be received. If pre-coding is used in such cases there is a danger that the recorded answers will be excessively stereotyped.

5.18 Arrangement of information on punched cards

No serious problems of card arrangement arise when the sampling units are the natural units of the population and the whole of the information on a unit can be coded on one card. The order in which the items are arranged on the card is immaterial in the Hollerith system. Consequently the order which is most convenient for punching may be adopted. Blank columns

between columns which are punched should be avoided by arranging all the blank columns at one end of the card.

It is generally advisable to leave a few blank columns in order to accommodate gang-punching of such items as index numbers and grouped classifications required in the course of the analysis. The number of blank columns likely to be required depends very greatly on the type of work. They may not be necessary at all in simple censuses for which the exact form of analysis can be prescribed at the outset. If in the course of the analysis it is found that additional columns are necessary, these can be made available by reproducing the cards, with the omission of items of information which are no longer required (or are not required in association with the additional columns).

If more than one card is needed to accommodate all the relevant information on a unit, it is necessary that items of information that require to be associated in the analysis should ultimately appear on the same card. Apart from the code number each item of information need only be punched on one card, the requisite items being transferred to other cards of the set by the reproducing punch. In certain cases entirely new cards may require to be constituted in this way, but in others sufficient blank columns may be left on the cards originally punched to accommodate the additional items. Convenience of punching should still be one of the prime considerations in the arrangement of the original cards. It is generally better to make an additional set of cards after punching than to separate information which falls in a natural punching sequence.

If the units which will form the basis of the analysis are not the sampling units, or if there is a hierarchy of units, the card arrangement presents more difficult problems. This situation is of fairly frequent occurrence. As an example we may consider suitable card arrangements for surveys of human populations in which the household is the sampling unit, and in which both households and individuals require to be treated as units in different parts of the analysis.

If the survey is a simple one, in which the whole of the information relating to the household can be accommodated in say 20 columns, and the whole of the information relating to each individual in say 10 columns, it will be possible to accommodate all the information relating to a household of up to five individuals on one card, leaving 10 columns blank for subsequent use. With this arrangement of the card, households of 6 to 10 individuals will require one trailer, 11 to 15 two trailers, etc.

This arrangement has the disadvantage that tabulations relating to individuals require five separate passages of the cards through the tabulator, with sorts between each passage, since individuals may appear in every one of five divisions of the card. This does not necessitate the passage of a larger total number of cards through the tabulator than if each individual were represented on a separate card, but it does result in the summaries being produced in five separate parts. These will then have to be combined by hand, or by punching

and tabulating summary cards. The total time of tabulation will also be somewhat increased owing to the increased number of printing and clearing cycles.

The alternative is to have a separate card for each individual. This will in any case be necessary if the amount of information relating to individuals is such that more than 10–15 columns are required per individual. It is usually necessary to have at least some of the information relating to the household reproduced on each individual card. If the household information requires say 40 columns and the individual information 30 columns, the household and the first individual in it can be punched on the first card. For the subsequent individuals only the code number of the household need be punched, the remainder of the household information being gang-punched subsequently. For convenience of punching, the code number should be so placed on the card that it is contiguous to the individual information. Each individual should also be allotted a serial number within the household in some orderly sequence, *e.g.* head of the household, wife, children by age, and other members by age. If still more space is required for the household information a separate card or cards will have to be given over to the household, with a selection of this information gang-punched on the cards for individuals.

The use of a separate card for each individual has one serious drawback. Although the whole of the information relating to a household and to all the individuals in it is recorded on the cards, it is impossible to classify households according to the collective characteristics of the individuals contained in them. We can of course pick out households containing one or more individuals having a given characteristic, *e.g.* we can select all households containing babies of under a year old by selecting the cards representing such babies. But we cannot, for example, classify the households according to number and age of children, unless this information has already been coded and recorded in summary form on the household part of the card.

In order to enable households having given collective characteristics to be picked out on the sorter, it is necessary for the whole of the relevant information concerning all the individuals in the household to be recorded on a single card. If, therefore, individuals in the same household are spread over more than one card we must construct a new set of household cards containing the relevant particulars of all individuals in the household. Some upper limit must be imposed on size of household, households of above this size being dealt with by hand where necessary. Thus with a 5-column field for household code number, and the reservation of 15 columns for subsequent recording of new classifications, it is possible to allot 5 columns to each of 12 individuals. The construction of such cards can be effected by reproducing on to the first set of 5 columns of the new cards the relevant columns of all individuals numbered 1, together with the household code numbers, followed by all individuals numbered 2 and so on. When the new household cards have been constructed they can be classified and machine coded according to their characteristics by sorting and gang-punching, using master cards.

Finally the new codes can be transferred to new household cards, together with the required household information.

This process is not necessarily simpler than the alternative of coding by hand from the original forms. In a small survey hand-coding will probably be more economical, particularly if the collective characteristics that require to be coded are known at the outset. Each case must be judged on its merits as it arises.

A further alternative, if the data on the original forms is not easily accessible, is to use the tabulator to list the relevant particulars of all individuals family by family, the coding being carried out by hand from this list.

5.19 General remarks on the planning of the computations

The preceding sections indicate the potentialities and limitations of the various methods of handling survey material. The methods of computation which will be most appropriate in given cases depend not only on the type of material, on the scale of the survey, and on the analysis required, but also on the equipment and personnel available. When punched card equipment is not readily accessible, for example, it may be better to use alternative methods, even though punched card equipment would otherwise be appropriate.

The extent to which it is advisable to carry out preliminary computations of index numbers, etc., on punched card equipment is also a matter which depends on the relative availability and cost of ordinary computing labour and punched card equipment. By suitable preliminary computations it is often possible materially to reduce the amount of work that has to be carried out on punched card machines. Thus, if only totals or means of multiple measurements taken on the same unit are required in the analysis it will clearly be better to obtain these before the cards are punched, punching only the results on the cards. On the other hand, such items as age at marriage can frequently be conveniently obtained from date of marriage and date of birth by gang-punching with interspersed master-cards. The question of the extent to which computations of this kind can be economically mechanized and carried out on punched card machines in any given circumstances is a very technical one, and the advice of an expert should always be sought before final decisions are taken.

In certain types of survey a good deal of preliminary computation is required on the forms before they can be abstracted and coded. Thus in the Survey of Fertilizer Practice (Section 4.23 and Example 6.19), which has been analysed on Cope-Chat cards, the applications of fertilizer are usually stated by the farmers in the form of so many hundredweights per acre of a given compound. From the dressing per acre and the chemical composition of the fertilizer it is necessary to work out the dressings per acre and the total dressings of the different plant nutrients, nitrogen, phosphate and potash. In the earlier surveys the weighting factors, which depended on the number of the fields and on the overall sampling fractions, were also introduced at this stage, both the

acres and the total dressings on these acreages being raised by these factors. In the later surveys a method of grouping the data according to the size of the weighting factor was adopted.

There is always a danger of over-mechanization in the analysis of survey material. The use of punched cards in particular can easily lead to a stereotyped and uncritical form of analysis. If surveys of the investigational type are analysed by punched card methods provision should always be made for further tabulations, etc., the need for which may become apparent when the results of the first tabulations have been examined in detail.

The degree to which an analysis can be planned at the outset depends very much on the type of survey. In a simple census type of survey the categories of information required may be determined at the outset by administrative requirements, and in this case the whole of the analysis can often be planned in advance. In surveys of the investigational type, however, it is only when the results have been examined and subjected to preliminary analysis that the most appropriate form for the final analysis will become apparent.

Even in surveys of the census type the information collected often forms a suitable basis for more detailed and critical statistical investigations. The possibility of such investigations should be considered at the outset when planning the coding of the material, so that the information will be available in accessible form if required. Items of information should not in general be omitted from the coding merely on the ground that they are not required for the primary analysis. The general aim should be to summarize the whole of the relevant information in coded form, so that should new needs arise or should the primary analysis indicate that further analysis is likely to be of value, the work can be undertaken without re-coding or the preparation of new cards.

5.20 Control of numerical accuracy in the analysis

The attainment of a high standard of accuracy in computational work is extremely difficult, and demands most careful organization of the checking procedures and scrupulous attention to detail at all stages. Moreover a reliable checking system can only be devised by careful study of the types of error which are likely to remain undetected.

Numerical work may be checked by repetition, by cross checks, or by using different methods of computation to arrive at the same results. Occasionally a comprehensive check which completely checks a large piece of computation is available, *e.g.* the solution of a set of linear equations must satisfy these equations. Reliable comprehensive checks and checks based on different methods of computation, however, are unfortunately rarely available in census and survey work.

Checking by repetition may consist of working over the same computation a second time, or carrying out the computations in duplicate. The original

and check computations may be carried out by the same or by different computers.

In a computation which is checked by working over the figures a second time the main causes of error may be classified as follows :—

- (1) Failure to check a value which is in error. (This includes partial failure, *e.g.* recalculation of the numerical value without checking the position of the decimal point.)
- (2) An identical error in both the original and check computation.
- (3) Different errors in the original and check computation which produce the same error in the next written or examined figure.
- (4) Failure to notice disagreement between the original and check computations when the original is in error.
- (5) Alteration of the original to agree with the check when the check is in error.
- (6) Failure to carry forward correctly corrections necessitated by a detected error.
- (7) Incorrect procedure in the original which is followed by the checker.

The danger of identical errors is obviously considerably greater when the same computer carries out both computations. Indeed at first sight it might appear that if a reasonably high standard of computing is attained, the chance of two computers making an identical error would be somewhat remote. There are, however, certain errors which are particularly common, such as, for example, the mis-reading of a badly written figure, incorrect location of the decimal point, the reversal of a pair of figures, *e.g.* 49,876 for 48,976, and duplication of the wrong figure, *e.g.* 74,496 for 74,996.

Duplicate computations, if properly carried out, *i.e.* not compared too frequently and corrected independently if an error is detected, will very greatly reduce the chances of most of the above types of error. Indeed a properly conducted set of duplicate computations done by different computers of reasonably high standard may be regarded as sufficiently accurate for almost all census and survey calculations. The checking of a single set of computations, however, even by different computers, cannot be expected to eliminate all errors, and if no other checks exist must be looked on as an unsatisfactory procedure for the more important computations.

Fortunately in many types of computation we do not have to rely solely on checks by repetition. A great deal of census and survey analysis is subject to cross checks of various kinds. Thus the counts relating to each of a number of classifications should all add up to the same total count. The same is true of totals of quantitative measurements. Indeed, where cross checks of this kind are not available it is often best to check a set of totals by calculating the grand total rather than by checking every individual total. The use of a grand total requires only a single comparison, which can consequently be made with some care. If all the individual totals have to be checked, there is serious danger that some discrepancy may be missed.

The use of cross checks in place of detail checks has a further advantage which is not so immediately apparent. If such a check fails to agree a good deal of re-computation is usually required to locate the error. This automatically tends to raise the standard of computation.

It is important to recognize which types of error will be detected by cross checks and which will not be so detected. If a total check is relied on, for example, all entries in a table are checked, but their locations in the table are not checked. Thus it is possible for quantities to be entered under the wrong headings. Such errors can be minimized by observing a standard order in all tables and always entering the values in the standard order.

One of the main functions of any checking system is to preserve a high standard in the computations. Very rigorous standards of work should be imposed. If more than a very few errors are found to exist a complete re-computation should be made. No erasures or fair copies must be permitted, and thorough inspection of all alterations must be made to see that errors are properly rectified. In large-scale routine work a record should be kept of the errors made by the different members of the staff. The supervisor must be ready at all times to resolve difficulties of procedure, otherwise the computers will undoubtedly attempt to resolve such difficulties amongst themselves, possibly incorrectly. A high standard of neatness must be insisted on. All figures must be legible and unambiguous not only to the writer but to others. This is particularly important in coding. Confusion between 6 and 0, and between X and Y, gives rise to many errors.

The coding and punching of the data of a large-scale census or survey presents its own organizational and checking problems. Even if a good deal of the information has to be coded it is advisable to record the coding on the questionnaire forms if possible, since transcription of pre-coded and numerical information is thereby avoided. If this is not possible a coding sheet may be used. This consists of an auxiliary printed form on which the information is entered in code, the form being so arranged that it is both convenient for the punch operators and for use in minor hand-analyses, if these are found to be required.

In certain cases the field investigators can be asked to code their own material at the end of each day's work. In general, however, this is not likely to be satisfactory, as it is difficult to preserve consistent standards of coding.

If the coding is at all difficult it is best to code one or a small group of items of information on a batch of forms at one time, rather than to code the whole of each form in turn. Whatever the detailed procedure adopted, however, it is essential for the supervisors to carry out adequate checks to ensure that correct and consistent standards are maintained.

In addition to routine checks it is often possible to impose checks of various kinds for gross errors and inconsistencies of coding. A special type of sorter, which picks out cards carrying a given code in a number of columns, can be used for this purpose.

Furthermore it is sometimes advisable to pick out extreme values and list the relevant particulars, so that it can be ascertained whether these values look reasonable.

In the above remarks we have been primarily concerned with the human element. It must not be assumed, however, that punched card equipment operates without error. The mechanism may fail in various ways, and such failure may be momentary only.

In order to devise adequate checks which at the same time are not unduly laborious, a full understanding of the mechanism is necessary. Thus, for example, if the control is operating on the columns on which the cards are sorted any mis-sort will be detected from the printed record, since the control will then break. Nevertheless sorting should always be given a preliminary check, either by passing a needle through the corresponding holes on each batch of cards, or, if the batches are small, by visual inspection, holding the batch up to a light.

The correctness of the totals provides adequate checks for much of the work on the tabulator, but it must be remembered that such totals are not fully checked by a single run of the cards through the machine, even if sub-totals are accumulated on one counter and the grand total on another, since there may be a faulty reading of a card. Equally, if a series of progressive totals are taken on a counter, the fact that the grand total is correct does not mean that all the progressive totals are correct, since the printing mechanism may have printed one of the numbers incorrectly.

The above remarks should not be taken to imply that any large number of errors are to be expected with punched-card equipment, but only that it must not be assumed that every sort and every printed figure is necessarily correct.

Whatever the methods of calculation, the final results of every analysis should be carefully scrutinized for apparent inconsistencies and irregularities, and any anomalous values should be thoroughly investigated.

Since the numerical material handled is itself in general subject to errors of various kinds, absolute numerical accuracy need not necessarily be attained at the early stages of the calculations. For this reason it is sometimes practicable to impose sample checks on such operations as punching and coding in large-scale work. If such checks are relied on, however, it is essential that steps are taken to prevent gross errors (Deming *et al.*, 1942, B).

Finally it should be emphasized that different types of work and different stages of the calculations demand very different standards of accuracy. A single misclassified card in a count, for example, will usually produce an entirely trivial error in the results. But the mispunching of a number representing a quantitative character, *e.g.* 610 for 010, may produce a serious error in the resultant mean or total of the class in which the unit falls. Errors in the final stages of the calculations are always likely to be more serious than those in the earlier stages. For this reason, in important work duplicate computations should be insisted on for the final stages, and these duplicates should themselves be used to check the typed or printed tables of the report.

5.21 The use of sampling in the statistical analysis

In certain cases it is possible to attain the necessary accuracy and at the same time to reduce the volume of numerical and machine work by analysing a sample of the available data. Such sampling may be applied to the data from a complete or sample census or survey.

At first sight the use of sampling in this manner appears illogical, since it might be argued that if the collection of the information on the whole of the population or on a large sample was justified, its inclusion in the analysis is also justified. This, however, is not always the case, since a complete census or a large sample may have been taken in order to furnish information on individual units or on small groups of units, while the further analysis may be required to elicit information which does not require to be broken down in detail. Moreover it does sometimes happen that excessively large samples are taken which can well be reduced before analysis.

As already mentioned, the analysis of a sample of the returns is also of use in providing preliminary results for a complete or sample census, even though the whole of the material will ultimately require analysis.

Furthermore, when a large sample has been taken for administrative purposes, supplementary analyses of the investigational type can often best be undertaken on a sub-sample of the original sample. The reduction in the total volume of material to be handled is of particular value in such analyses, since they often require the application of relatively complicated statistical processes. Special points which emerge and on which a higher accuracy is desired can be re-tabulated subsequently by using the whole or a larger sub-sample of the material.

The actual technique of obtaining a sample suitable for analysis is usually relatively simple. For many purposes a systematic sample of every q th return is all that is required. In some cases, however, the use of a variable sampling fraction is advisable. This is particularly the case in the analysis of census returns referring to economic institutions, factories, farms, etc., since these are usually of very variable size.

An example of an analysis of this type is provided by the National Farm Survey of England and Wales (Ministry of Agriculture, 1944, G), which covered all holdings in England and Wales of over 5 acres. Sampling was not used in the survey because records were required for each individual farm, both for administrative purposes and for detailed studies of small areas. A map of the boundaries of each farm, for example, was one item of information which was collected.

For the purpose of obtaining a general summary of the results by counties, types of farming, etc., the analysis of the whole of the material was unnecessary. The holdings were therefore divided into size-groups and a systematic sample stratified for counties and size-groups was taken, using a variable sampling fraction for size-groups. The sampling fractions, and numbers of holdings in the population and in the sample, are shown in Table 5.21.

TABLE 5.21—ANALYSIS OF THE NATIONAL FARM SURVEY :
CONSTITUTION OF SAMPLE

Size-group (acres)	Average size (acres)	No. of holdings	Sampling fraction (per cent.)	No. of holdings in sample
5- 25	12	101,450	5	5,072
25-100	55	111,360	10	11,136
100-300	165	65,210	25	16,302
300-700	413	11,150	50	5,575
Over 700	1,035	1,430	100	1,430
		290,600	(13.6)	39,515

Had a uniform sampling fraction been used in place of a variable sampling fraction, a sample over twice as large would have been required to give results of the same accuracy on such items as the percentage of land under different systems of tenure. By the use of a variable sampling fraction results of ample accuracy were obtained from an analysis covering only one-seventh of all the holdings. This not only considerably reduced the amount of coding and machine work, but also enabled work to proceed as soon as the information for the sample farms had been assembled and abstracted. In consequence it was possible to make the results of the analysis available a year or two sooner than would have been the case had the whole of the material had to be abstracted before analysis.

5.22 Adjustment of the results to compensate for defects in the sample

When the sampling procedure is defective in one respect or another, attempts are sometimes made to adjust the results in order to compensate for the defects. Thus it may happen that owing to defects in the selection of the sample or in the collection of the information, different classes of the population are found to be represented in incorrect proportions in the final sample. In such cases it is possible to adjust the results by weighting the different classes in such a manner as to compensate for the errors in the proportions.

This procedure must be clearly distinguished from the procedure of stratification after selection mentioned in Section 3.3. The validity of the latter procedure depends on the fact that the sample as a whole is random and therefore the selection from within strata is also random. If the proportions in the different classes are different because of defects in the sampling procedure, however, it is most unlikely that the selection from within these classifications will be fully random. Any adjustment of the type envisaged, therefore, although it may somewhat improve matters, must not be expected to eliminate by any means the whole of the defects.

Stratification after selection is a special case of the use of supplementary information of all kinds. Such adjustments, whether planned at the outset or decided on subsequently after examination of the data, are quite justified. The essential difference between these adjustments and between adjustments of the same type made in order to compensate for defects in the sampling procedure is that in the former the selection is random, except for permissible restrictions, whereas in the latter it may be biased in various ways.

In general, if the sampling procedure is defective it is best to report the results obtained without adjustment. At the same time data should be given indicating, so far as is possible, the deviations of the sample from the expected distributions. Thus if the proportions in the different classes of a classification are known for the population these may be presented alongside the parallel classification of the sample. Similarly the sample means of quantities for which the population means are known may be presented for comparison. Occasionally an adjustment of some of the more important values derived from the sample may be considered worth while, but in such cases the unadjusted results should also be presented.

The above remarks apply primarily to samples for which the sampling procedure is markedly defective. In cases in which there are slight defects, such as a minor degree of non-response, the application of some small adjustment, if this appears necessary, is more justified. If such adjustments are made, however, the fact should be clearly stated and their magnitude should be indicated.

The simplest way of dealing with non-response is to regard the non-respondents as similar to the remainder of the sample, *i.e.* to treat the sample as if it were a sample on a smaller number of units. With a stratified sample, the non-respondents in each stratum can be treated as the equivalent of respondents in that stratum. Alternatively some other appropriate classification can be used, as indicated above.

If follow-up methods have been used and there has been a good response to the follow-up, initial non-respondents who subsequently respond can be treated as a sub-sample of all initial non-respondents and weighted accordingly. It is clear that if there is any difference between respondents and non-respondents the final non-respondents may be expected to be more like the initial non-respondents than the general population. This procedure was first, so far as I know, suggested by Professor D. V. Glass, and was used by him in the analysis of the Family Census (Section 4.10).

In this survey those who failed to provide the enumerator with the required information were sent a letter further explaining the purposes of the survey and requesting that the form be sent direct to the Royal Commission. Of the 230,000 initial non-respondents (*i.e.* 17 per cent. of the whole sample), 50,000 responded to this appeal. This 50,000 therefore constituted a sample (though a non-random one) of the 230,000, and the first 12,000 of the 50,000 replies were combined with the remainder of the sample with a weight of 230/12. This procedure was found to give overall birth-rates which corresponded

very closely to those already known from other sources, whereas the original sample gave birth-rates which were substantially too high, owing to the fact that the majority of the initial non-respondents were women with few or no children.

5.23 Critical analysis of survey data

At the outset a clear distinction must be made between the types of deduction that can be made with certainty from survey data and the types that are speculative. If in a nutrition survey, for example, we find that children of large families are worse fed than children of small families we can draw the definite conclusion that size of family is associated with malnutrition of the children, and we can give quantitative estimates of the degree of malnutrition actually existing amongst children of families of different sizes. We cannot, however, assert with certainty that size of family is the cause of this malnutrition, though the fact that in large families the income per head is automatically less if there is a fixed total income would lead us to expect an underlying causal relationship.

Even in situations where a definite causal relationship is known to exist, deductions as to the magnitudes of the effects of given factors can never be made with certainty from survey data. We may, for instance, find that fields receiving fertilizers give higher yields per acre than fields without fertilizers. Yet we cannot attribute the observed differences solely to differences in fertilizers. The farmers using the fertilizers may be farming better land, they may be growing higher-yielding varieties, and they may be carrying out their farming operations with greater skill.

Clearly definable extraneous factors which may influence the estimates of the effects of other factors can be determined in the course of the survey. Under certain circumstances the disturbance due to them can be eliminated by methods of analysis which will be outlined in this and the following section. But there will always be other undetermined and possibly unascertainable factors which cannot be taken into account.

In order to determine with certainty the magnitude in the causal sense of the effect of any given factor, experiments must be undertaken. Surveys cannot be regarded as satisfactory substitutes for experiments. Nevertheless they are of value in situations in which experiments are difficult or impossible, though in such cases all conclusions must be tentative. They are also of value as a preliminary to experimental work, since they frequently indicate the factors that are likely to be most worth investigation.

If, however, survey data are to be effectively used for either of these two purposes it is important to have means of eliminating the effects of extraneous factors in so far as this is possible.

A simple example will illustrate the problem involved. Table 5.23.a gives the numbers of fields, totals and means of yields per acre of a sample of 901 potato fields classified according to (a) the five regions into which the country was divided, and (b) the five varieties included in the survey. These

results were obtained in the course of an investigation into the blackening of potatoes on cooking, the data on yields being collected from the farmers when the samples were taken, together with a considerable amount of information on fertilizers, cultural practices, etc. Approximately 180 fields were selected in each region. The selection within regions was not strictly random, but can be regarded as substantially so for the purpose of the present discussion. The sample was confined to the five named varieties, but was not stratified by varieties.

From the results of Table 5.23.a it is apparent that the mean yield is highest for the Scottish region, and is also higher for the Northern region than for the remaining regions. There are, however, even larger varietal differences in yield. Consequently if the varieties were grown in different proportions in the different regions the regional differences are likely to be influenced by varietal differences.

To examine this point it is necessary to construct the two-way classification, regions $\times$ varieties. This is shown in Table 5.23.b. The values of Table 5.23.a appear as marginal totals in this table.

TABLE 5.23.a—POTATO SURVEY: NUMBERS OF FIELDS AND TOTALS AND MEANS OF THE YIELDS PER ACRE (TONS)

(a) *Classified by regions*

	No.	Total	Mean
Scotland . . .	174	1,482	8.52
North . . .	177	1,425	8.05
E. Midlands . .	189	1,415	7.49
South . . .	182	1,324	7.27
West . . .	179	1,368	7.64
ALL . . .	901	7,014	7.78

(b) *Classified by varieties*

	No.	Total	Mean
Majestic . . .	393	3,292	8.38
King Edward . .	250	1,563	6.25
Great Scot . . .	56	461	8.23
Arran Banner . .	84	766	9.12
Kerr's Pink . . .	118	932	7.90
ALL . . .	901	7,014	7.78

TABLE 5.23.b—POTATO SURVEY: TWO-WAY CLASSIFICATION OF THE DATA BY REGIONS AND VARIETIES

Numbers of fields

	Scot.	North	E. Mid.	South	West	Total
Majestic . .	37	75	104	101	76	393
King Edward .	42	14	85	66	43	250
Great Scot .	18	14	—	6	18	56
Arran Banner .	8	38	—	9	29	84
Kerr's Pink .	69	36	—	—	13	118
TOTAL .	174	177	189	182	179	901

Totals of yields per acre (tons)

	Scot.	North	E. Mid.	South	West	Total
Majestic . .	350	614	876	823	629	3,292
King Edward .	321	80	539	387	236	1,563
Great Scot .	166	106	—	49	140	461
Arran Banner .	73	351	—	65	277	766
Kerr's Pink .	572	274	—	—	86	932
TOTAL .	1,482	1,425	1,415	1,324	1,368	7,014

Means of yields per acre (tons)

	Scot.	North	E. Mid.	South	West	All
Majestic . .	9.46	8.19	8.42	8.15	8.27	8.38
King Edward .	7.65	5.71	6.34	5.87	5.49	6.25
Great Scot .	9.22	7.57	—	8.17	7.78	8.23
Arran Banner .	9.12	9.24	—	7.22	9.54	9.12
Kerr's Pink .	8.29	7.61	—	—	6.61	7.90
ALL . .	8.52	8.05	7.49	7.27	7.64	7.78

From the first part of this table (numbers of fields) it is apparent that the distribution of varieties is by no means the same for all regions. In particular very little King Edward, which yields about 2 tons per acre less than the other varieties, was grown in the Northern region. The relatively high yield of the Northern region is therefore accounted for, in part at least, by varietal differences.

To make an estimate of what the differences between regions would be if the proportions of the different varieties were the same in all regions we can compare the regional means of the individual varieties. These are given in the last part of Table 5.23.b. Inspection shows that Scotland gives higher yields than every other region for all varieties except Arran Banner, of which there are only 8 fields in Scotland. The Northern region, on the other hand, does not show any consistent differences from the other English regions.

Inspection of this kind may in certain cases be all that is necessary. Frequently, however, quantitative estimates of the differences attributable to one classification when freed from the effects of a second classification are required. Such estimates may be obtained in various ways, depending on the nature of the table and which differences are of interest.

(1) *Unweighted means of sub-class means*

If all the sub-class means are of adequate accuracy the marginal unweighted means of these means can be taken. These unweighted means will give estimates of the differences attributable to either classification when the units of each class are equally divided between the classes of the other classification.

This method cannot be applied to the whole of Table 5.23.a because of the unoccupied cells, but it can be applied to the parts of the table represented by all varieties in the Scottish, Northern and Western regions, or all regions for varieties Majestic and King Edward. The results are shown in Table 5.23.c. For comparative purposes each set of means has been adjusted by adding a constant amount so that the mean of the set is equal to the general mean. The similarity of the Northern region with the other English regions is confirmed. The yield of Kerr's Pink has also been reduced relative to the other varieties. This is a consequence of the high proportion of Kerr's Pink in Scotland.

(2) *Standardization of proportions by weighted means of sub-class means*

This is similar to Method 1. The weighted means of columns (or rows) of the table of sub-class means are taken, with weights roughly in proportion to the numbers in each row (or column) class in the whole sample.

An example of this method is given in Example 6.8. It is not applicable in full to Table 5.23.b on account of unoccupied cells.

It should be noted that somewhat different quantities are estimated by the two methods. Failure to recognize this fact sometimes causes a certain amount of confusion. If applied to a varieties $\times$ regions table, for example, Method 1

estimates the differences between regions that would occur in the hypothetical situation in which equal numbers of fields of all varieties were grown in each region. Method 2 gives estimates appropriate to the hypothetical situation in which the different varieties are grown in the same proportions in all regions, these proportions being equal to the average proportions for the whole country. Only if the differences between the different varieties are the same for all regions will the estimated differences be the same.

Method 2 has two advantages over Method 1. It is in general more accurate, since greater weight is on the average given to the cells containing the greater numbers of units. It also gives estimates which refer to a hypothetical situation more in conformity with that actually existing.

TABLE 5.23.c—POTATO SURVEY: UNWEIGHTED MEANS OF SUB-CLASS MEANS
(a) OMITTING E. MIDLANDS AND SOUTHERN REGIONS, (b) OMITTING GREAT SCOT, ARRAN BANNER AND KERR'S PINK

	Unadjusted means		Means adjusted to sample mean	
	(a)	(b)	(a)	(b)
Scotland . . .	8.75	8.56	8.55	8.98
North . . .	7.66	6.95	7.46	7.37
E. Midlands . . .	—	7.38	—	7.80
South . . .	—	7.01	—	7.43
West . . .	7.54	6.88	7.34	7.30
Mean . . .	7.98	7.36	7.78	7.78
Majestic . . .	8.64	8.50	8.44	8.92
King Edward . . .	6.28	6.21	6.08	6.63
Great Scot . . .	8.19	—	7.99	—
Arran Banner . . .	9.30	—	9.10	—
Kerr's Pink . . .	7.50	—	7.30	—
Mean . . .	7.98	7.36	7.78	7.78

It will be recognized that the marginal means of the sub-class means, whether weighted or unweighted, do not contain the whole of the information. If variety *P* yields more than variety *Q* in one region and less in another, for example, this fact can only be established from the sub-class means. Under such circumstances any comparison of the regions based on equalization of the proportions of the varieties represents an over-simplification of the real situation.

Neither Method 1 nor Method 2 can be applied to the whole of tables in which there are blank cells. Even if there are no blank cells neither method will be very satisfactory when there are certain cells which contain so few units that the corresponding cell means are very inaccurate. Thus in the present example the relatively large difference between Scotland and the Northern region in column (b) of Table 5.23.c, in contrast to column (a),

is in part due to the fact that Arran Banner has given a larger yield in the Northern region than in Scotland. This may well be due to sampling errors, since there are only 8 fields of Arran Banner in Scotland.

There are two further relatively simple methods which are of value in such circumstances.

(3) *Weighted means of differences of sub-class means*

If only two classes (cross-classified by a second classification into a number of sub-classes) are to be compared, a weighted mean of the differences of each pair of sub-classes can be taken. Maximum accuracy will be attained when the weights are inversely proportional to the squares of the standard errors of these differences.

With independent samples in each sub-class the square of the standard error of a difference is equal to the sum of the squares of the standard errors of the two means (Section 7.5). Under certain circumstances, which will be apparent from a study of Chapter 7, and in particular when the selection from within sub-classes is effectively random and the standard deviation per unit is constant, the standard errors are inversely proportional to the square roots of the numbers of units in the sub-classes (Section 7.1). In this case the reciprocals of the weights must be taken proportional to the sums of the reciprocals of the pairs of sub-class numbers, *i.e.* if n_1 and n_2 are a pair of sub-class numbers the weight can be taken equal to w , where

$$\frac{1}{w} = \frac{1}{n_1} + \frac{1}{n_2}$$

The calculations are shown in Table 5.23.d. The weight for Majestic, for example, is given by $1/w = 1/37 + 1/75$. Weights may be taken to the nearest

TABLE 5.23.d—POTATO SURVEY: ESTIMATE OF DIFFERENCE OF SCOTTISH AND NORTHERN REGIONS FROM WEIGHTED MEAN OF VARIETAL DIFFERENCES

	Difference z	Weight w	wz
Majestic . . .	+ 1.27	25	+ 31.75
King Edward . .	+ 1.94	10	+ 19.40
Great Scot . . .	+ 1.65	8	+ 13.20
Arran Banner . .	— 0.12	7	— 0.84
Kerr's Pink . . .	+ 0.68	24	+ 16.32
Total	—	74	+ 79.83
Weighted Mean . .	+ 1.08		

whole number. They can be rapidly calculated from a table of reciprocals or on a slide rule. The weighted mean, + 1.08, is obtained by dividing the total of wz by the total of w .

(4) *Pooling of classes*

In cases in which inspection or use of the previous methods indicates that the differences between certain of the classes are small, such classes can be pooled. This often eliminates blanks and very small numbers in the sub-class table.

In the present example inspection indicates that there is little difference between the four English regions. This is confirmed by the means in Table 5.23.c. These regions may therefore be pooled. This pooling will permit a better estimate of the differences between the last three varieties than that given by Table 5.23.c. After pooling Scotland can be included at $\frac{1}{5}$ weight, following Method 2.

The calculations are shown in Table 5.23.e. It will be noted that the pooling can be effected by adding the numbers of fields and totals of yields for the four English regions from Table 5.23.b.

TABLE 5.23.e—POTATO SURVEY: EFFECT OF POOLING ENGLISH REGIONS

	English regions (pooled)			Scotland : Mean	Weighted mean
	No.	Total	Mean		
Majestic . . .	356	2,942	8.26	9.46	8.50
King Edward . .	208	1,242	5.97	7.65	6.31
Great Scot . . .	38	295	7.76	9.22	8.05
Arran Banner . .	76	693	9.12	9.12	9.12
Kerr's Pink . . .	49	360	7.35	8.29	7.54
ALL	727	5,532	7.61	8.52	7.79
Weight			4	1	

5.24 Method of fitting constants

If there are more than two classes between which differences are required, Method 3 can be used to compare each pair separately. It will not, however, give a consistent set of estimates, *i.e.* the sum of the estimated differences between *A* and *B* and between *B* and *C* will not exactly equal that between *A* and *C*. This is a reflection of the fact that Method 3, though fully efficient if there are only two classes in the table, is not fully efficient (in the statistical sense) when there are more than two classes. In this case, in addition to direct comparisons between *A* and *B*, indirect comparisons of *A* with *C* and *C* with *B*, etc., can be made. When the sub-class frequencies are not proportionate these will contribute some additional information. To take an extreme case, if variety

P occurs in regions *A* and *C* only, and variety *Q* in regions *B* and *C* only, comparisons between regions *A* and *B* with differences between varieties eliminated can only be made by indirect comparisons involving region *C*.

Estimates of maximum accuracy, which, as might be expected, are also consistent, can be obtained by fitting constants by the method of least squares (Yates, 1934, A ; Snedecor, 1934, A ; Snedecor and Cox, 1935, A). Snedecor has drawn attention to the fact that if the numbers of units in the different sub-classes are nearly proportionate, Method 2 can be used without appreciable loss of information in place of the more laborious method of fitting constants.

Stevens (1948, A) has given a simple arithmetical method of obtaining the values of the estimates derived by fitting constants. The procedure, which is one of successive approximation, is illustrated for the data of our example in Table 5.24.

TABLE 5.24—POTATO SURVEY: ESTIMATION OF VARIETAL AND REGIONAL DIFFERENCES BY FITTING CONSTANTS

		Sc.	N.	E.M.	S.	W.	Starting values and corrections			Final values	
	901	174	177	189	182	179	(1)	(2)	(3)	(4)	(5)
Maj.	393	37	75	104	101	76	+ .60	+ .10	+ .02	+ .73	8.51
K.E.	250	42	14	85	66	43	-1.53	.00	+ .02	-1.51	6.27
G.S.	56	18	14	—	6	18	+ .45	-.08	-.03	+ .34	8.12
A.B.	84	8	38	—	9	29	+1.34	+ .16	.00	+1.50	9.28
K.P.	118	69	36	—	—	13	+ .12	-.39	-.08	- .35	7.43
Starting values and corrections	(1)	+ .74	+ .27	-.29	-.51	-.14					
	(2)	+ .09	-.48	+ .36	+ .14	-.16					
	(3)	+ .83	-.21	+ .07	-.37	-.30					
	(4)	+ .13	+ .01	-.06	-.06	-.03					
	(5)	+ .03	+ .01	-.02	-.02	.00					
Final values	(6)	+ .99	-.19	-.01	-.45	-.33					
	(7)	8.77	7.59	7.77	7.33	7.45					

The number of fields in each sub-class, reproduced from Table 5.23.b, is shown in the body of the table, with marginal totals above and to the left. Column 1 and row 1 give the deviations of the varietal and regional mean yields from the general mean 7.78. These are obtained from Table 5.23.a or 5.23.b. Thus, for Majestic, $8.38 - 7.78 = + 0.60$. These data are all that are necessary for the estimation process.

The approximation should be started with the set of deviations which show the biggest differences, in this case the varietal deviations. We first

calculate what the regional deviations would be with these varietal deviations if there were no regional differences. These regional deviations are shown *with signs reversed* in row 2. To obtain them the numbers of fields in each column are multiplied in turn by the deviations in column 1, summed and divided by the total number of fields in each column. Thus for Scotland the deviation is

$$(+.60 \times 37 - 1.53 \times 42 + .45 \times 18 + 1.34 \times 8 + .12 \times 69)/174 \\ = -14.96/174 = -.09$$

and similarly for the other columns. It is best to record the sums of products, -14.96 , etc., before division, as this provides a check against minor errors, the total of these sums of products being equal to the sum of the products of column 1 and the total column of numbers of fields. This sum, $+5.22$, differs from zero only because of rounding-off errors.

Since the observed deviation for Scotland is $+.74$, and the expected deviation if there were no regional differences is $-.09$, a first estimate of the true regional deviation for Scotland with varietal differences eliminated is $+.74 - (-.09) = +.83$. These estimates are shown in row 3, which is obtained by adding row 2 to row 1.

The varietal deviations in column 1, however, are themselves affected by regional differences. These may now be corrected for by the same process, using the estimates of the regional deviations just obtained. To do this the values of row 3 are multiplied in turn by the numbers of fields in each row, summed and divided by the total number of fields in the row. Thus for Majestic we obtain, after reversal of sign, the correction $+.10$. These corrections are shown in column 2. The checks operate as before.

The corrections in column 2 could now be added to the values of column 1 to give second approximations to the varietal differences. It is simpler, however, to use the corrections themselves to calculate corrections to the estimated regional differences of row 3. These are shown in row 4. The same procedure of calculation is followed, and the signs are reversed as before.

The values of row 4 are then used to give a second set of varietal corrections, which are shown in column 3, and these in turn are used to give a third set of regional corrections, which are shown in row 5.

The process may be stopped at this point, since the corrections are now so small that the next set will be negligible. Columns 1, 2 and 3 and rows 3, 4 and 5 are therefore summed to give the final estimates of the deviations, shown in column 4 and row 6. These deviations may then be added to the general mean 7.78 to give final estimates of the varietal means freed from regional differences (column 5), and of regional means freed from varietal differences (row 7). Final checks are obtained by forming the sums of products of these means with the total numbers of fields and dividing by the grand total 901 . In each case the general mean 7.78 should be obtained.

It will be seen that the final values do not differ greatly from the corresponding values of Tables 5.23.c, 5.23.d and 5.23.e, but they do differ substantially

from the values of Table 5.23.a. The differences between Scotland and the Northern region, for example, are as follows:—

Regional means over all varieties (Table 5.23.a)	..	..	0.47
Unweighted means of sub-class means (Table 5.23.c)	{ (a) ..		1.09
	{ (b) ..		1.61
Weighted differences (Table 5.23.d)	..	..	1.08
Fitting constants (Table 5.24)	..	..	1.18

Equally the varietal means given in Table 5.23.e are very close to those of Table 5.24.

This demonstrates that the more direct methods, used with judgment, are capable of giving satisfactory estimates. As is shown in Example 7.5, where the errors of these estimates are discussed, the third of the above regional differences, viz. 1.61, is decidedly less accurate than the others, since the information provided by the last three varieties is not taken into account. On the other hand, the agreement of the above estimates must not be taken as providing any indication of their real accuracy. Since all the estimates are based on the same data, any disagreement is primarily a reflection of the effects of the various approximations on the efficiency of the estimation process.

There are some further general points in connection with the method of fitting constants which should be noted.

The method will provide efficient estimates of the differences due to one classification, freed from the effects of the other classification, if the true differences are the same for all classes of the second classification. In the terminology of the design of experiments, this is equivalent to the non-existence of *interactions*. If the true differences vary markedly, the method is inappropriate. Instead the individual differences should be considered or Method 1 or Method 2 should be used. (See sub-section (2) above.)

As mentioned above, if one of the classifications consists of two classes only, Method 3 is fully efficient for estimating the difference between these two classes. If estimates of the differences between the classes of the second classification are required, these may be derived by adjusting the class means in the manner followed in Table 5.24, assigning deviations of plus and minus half the difference to the two classes of the first classification. The method is in this case exact, and therefore no further approximations are required.

The method can be extended to multiple classifications having three or more sets of classes. In this case the data required are the general means for each main classification together with the two-way tables of numbers of units corresponding to each pair of classifications. The numbers of units in the individual cells of the three- or more-way table are *not* required. If there are only two classifications the general means for each main classification, together with the numbers of units in the individual sub-classes, are required.

The reporting of sub-class numbers of the relevant pairs of classifications should therefore be considered even in cases in which the reporting of the

separate sub-class means or the calculation of adjusted estimates is not considered worth while. If these numbers are available the various inter-relations may then be studied subsequently without further reference to the original material.

5.25 Preparation of reports

When the analysis of a census or survey has been completed it is usually necessary to embody the results in a report. In addition to the presentation of the numerical results in the form of tables and graphs, some discussion and interpretation is also required.

The lines to be followed in the preparation of such reports vary greatly according to the nature of the material and the purposes of the report, but are in general similar for sample censuses and surveys and for complete censuses and surveys. We therefore do not propose to discuss them here.

There are, however, certain matters which should be reported on in a sample census or survey which do not arise (or are of less importance) in a complete census. These matters have been covered by the memorandum (already referred to in Section 1.6) prepared by the United Nations Sub-Commission on Statistical Sampling, entitled *Recommendations concerning the Preparation of Reports on Sampling Surveys*. The recommendations* are as follows:—

(1) General description of the survey

The general description of the survey should include information on the following points. Some of these will require fuller treatment in the more detailed technical sections of the report.

- (a) Statement of purposes of the survey. A general indication should be given of the purposes of the survey and the ways in which it had been expected that the results would be utilized.
- (b) Description of the material covered. An exact description should be given of the geographical region and the categories of material covered by the survey. In a survey of a human population, for example, it is necessary to specify whether such categories as hotel residents, institutions (e.g. boarding houses, sanatoriums), vagrants, military personnel, were included. The reporter should guard against any possible misapprehension regarding the coverage of the survey.
- (c) Nature of the information collected. This should be reported in considerable detail, including a statement of items of information collected but not reported on. The inclusion of copies of the schedules and relevant parts of the instructions used in the survey (including special rules for coding and classifying) is often of value. If this is impracticable, it may be possible to make available a limited number of copies which may be obtained on request.

* The introduction to the memorandum is omitted.

- (d) Method of collecting the data. Whether by interviewers, investigators, mail, etc.
- (e) Sampling method. An indication should be given in general terms of the type of sampling adopted, the size of the sample, the proportion it forms of the material covered, and arrangements for follow-ups, if any, in cases of non-response.
- (f) Accuracy. A general indication of the accuracy attained should be given.
- (g) Repetition. State whether the survey is an isolated one undertaken without intention of repetition, or is one of a series of similar surveys.
- (h) Point or period [of time]. Point or period of time to which the data refer.
- (i) Date and duration. The starting date and period taken for the field work.
- (j) Cost. An indication should be given of the cost of the survey, under such headings as preliminary work, field investigations, analysis, etc.
- (k) Responsibility. The name of the organization sponsoring the survey and of the one responsible for conducting it.
- (l) References. References should be given to any published reports or papers.

(2) *Design of the survey*

The [sampling] design of the survey should be carefully specified.*

(3) *Method of selecting sample-units†*

The reporter should describe the procedure used in selecting sample-units, and if it is not a random selection he should indicate the evidence on which he relies for adopting an alternative procedure. Purposive selection and quota sampling cannot be regarded as equivalent to random sampling.

(4) *Personnel and equipment*

It is desirable to give an account of the organization of the personnel employed in collecting, processing and tabulating the primary data, together with information regarding their previous training and experience. Arrangements for training, inspection, supervision, and methods of processing data should be explained, as also should methods of checking the accuracy both of the primary data and of the processing. A brief mention may be made of the equipment used in processing the data.

* A section on terminology follows.

† The first paragraph, defining random and systematic processes of selection, is omitted.

The critical observations of the technicians in regard to any part of the survey should be given. These observations will help others to improve their operations.

(5) *Costs*

An important reason for the use of sampling (instead of complete enumeration) is lower cost. Information on costs is therefore of great interest. Costs should be classified so far as possible under such heads as preparation (showing separately the cost of pilot studies), field work, supervision, processing, analysis, and overhead costs. In addition, labour costs in man-weeks of different grades of staff, and also time required for interview and journey time and transport costs between interviews, should be given. The compilation of such information, although often inconvenient, is usually worth undertaking as it may suggest substantial economies in the planning of future surveys. Efficient design demands a knowledge of the various components of cost, as well as of the components of variance.

(6) *Accuracy of the survey*

- (a) Precision as indicated by the random sampling errors deducible from the survey. Standard deviations of sample-units should be given in addition to such standard errors (of means, totals, etc.) as are of interest. The process of deducing these estimates of error should be made entirely clear. This process will depend intimately on the design of the sample survey. An analysis of the variances of the sampling-units into such components as appear to be of interest for the planning of future surveys is also of great value.
- (b) Degree of agreement observed between independent investigators covering the same material. Such comparison will be possible only when interpenetrating samples have been used, or checks have been imposed on part of the survey. It is only by these means that the survey can provide an objective test of possible personal equations (differential bias among the investigators).
- (c) Other non-sampling errors. (i) Errors which are common to all investigators, and indeed any constant component of error (or "bias") in the recorded information, will not be included in the estimates of the random sampling errors deducible from the survey results. (ii) Another source of error of the same type is that due to observation of quantities which do not correspond exactly to the quantities of which estimates are required: in a crop-cutting survey, for example, the yields of the sample plots give estimates of the amount of grain, etc., in the standing crop, whereas the final yield will be affected by losses at harvest. (iii) The possible effects of such errors on the accuracy of the results, and of incompleteness in the recorded information (*e.g.* non-response, lack of records, whether covering the whole of the

survey or particular areas or categories of the material), should therefore be fully discussed. (iv) Any special checks instituted to control and determine the magnitude of these errors should be described, and the results reported.

- (d) Accuracy, completeness and adequacy of the frame. The accuracy of the frame can and should be checked and corrected automatically in the course of the enquiry, and such checks afford useful guidance for the future. Its completeness and adequacy cannot be judged by internal evidence alone. Thus complete omission of a geographic region or the complete or partial omission of any particular class of the material intended to be covered cannot be discovered by the enquiry itself, and auxiliary investigations have often to be made. These should be put on record, indicating the extent of inaccuracy which may be ascribable to such defects.
- (e) Comparison with other sources of information. Every reasonable effort should be made to provide outside comparisons with other sources of information. Such comparisons should be reported along with the other results, and the significant differences should be discussed. The object of this is not to throw light on the sampling error—since a well-designed survey provides adequate internal estimates of such errors—but rather to gain knowledge of biases, and other non-random errors.
- (f) Efficiency. The results of a survey often provide information which enables investigations to be made on the efficiency of the sampling designs, in relation to other sampling designs which might have been used in the survey. The results of any such investigations should be reported. To be fully relevant the relative costs of the different sampling methods must be taken into account when assessing the relative efficiency of different designs and intensities of sampling. Such an investigation can be extended to consideration of the relation between the cost of carrying out surveys of different levels of accuracy and the losses resulting from errors in the estimates provided. This provides a basis for determining whether the survey was fully adequate for its purpose, or whether future surveys should be planned to give results of higher or lower accuracy.

ESTIMATION OF THE POPULATION VALUES

6.1 Possibility of alternative estimates

In this chapter we shall deal with the derivation of estimates of the population values from the numerical results obtained in the sampling. A simple example of such an estimate is provided by the arithmetic mean of the sample values of a random sample. It is well known that this mean provides an estimate of the mean of the population from which the sample was drawn, though it will not, owing to sampling errors, be exactly equal to the mean of the population.

The arithmetic mean of the sample values is not the only possible estimate of the population mean. We might, for instance, take the median, *i.e.* the central value, or the geometric mean, *i.e.* the antilogarithm of the mean of the logarithms of the sample values, or even the mean of the highest and lowest values in the sample.

In addition to estimates such as the mean and the median, which can be derived from a given set of values independently of any supplementary information associated with these values, there are further alternative estimates which can be derived by taking account of such supplementary information as is available, either qualitative or quantitative. Thus, as has been mentioned in Section 3.3, if the numbers of units from the whole population falling in the different strata of some stratification are known, a random sample can be adjusted so that the different strata are represented in their correct proportions. Similarly, supplementary information on a quantitative character can be used in various ways to provide estimates which will in general be more accurate than the simpler estimates which do not utilize this information.

In deciding which is the best estimate for any given type of sampling three different criteria have to be considered. These are, absence of bias, accuracy (or, as it is technically known, efficiency), and computational convenience. In the case of a random sample, if the population values are normally distributed—the meaning of this term is explained in Chapter 7—the arithmetic mean will provide an estimate which is both free from bias and, apart from supplementary information, of maximum accuracy. It is also sufficiently simple computationally for practical use. More important, the mean will remain an unbiased estimate of the population mean whatever the form of the distribution of the population values, though it will not necessarily be the most accurate estimate that could be devised. The mean also has the incidental advantage that the sampling errors to which it is subject can be relatively easily assessed, and are not greatly dependent on the form of the distribution of the population values.

In this book we do not propose to do more than list what appear to be the most useful estimates for any given type of sampling, and give examples

of the computations involved. Any discussion of questions of bias and relative efficiency requires advanced mathematical statistical theory. In general the recommended estimates are free from any important source of bias ; where this is not the case the circumstances in which bias can arise are indicated.

Estimates are required not only for the whole population, but frequently also for the different parts of it which constitute domains of study. The formulæ of estimation which are applicable to the whole population are in general applicable also to the separate domains, and need not be discussed separately. In certain cases adjustments which can be applied to the population estimates cannot be applied to the estimates for the different domains. Thus if the population mean of a supplementary variate is known, but not the means for the different domains of study, adjustment by means of supplementary information can be applied only to the estimates of the whole population. In like manner the gain in accuracy due to stratification is sometimes different for the population and domain estimates. If the domains cut across the strata, for example, the errors of the domain estimates may only be slightly reduced by the stratification.

6.2 Notation

It is important, both in discussion of the problem of estimation and in the mathematical formulæ, to make a clear distinction between estimates of the population values and the population values themselves. The present convention in mathematical statistics is to denote the population parameters by Greek letters and the corresponding estimates by the corresponding Latin letters. This convention, however, is difficult to apply consistently, and is in any case more appropriate for infinite hypothetical populations than for the finite populations met with in sampling. In the present manual we have for the most part adopted the convention of denoting the population values by bold type, the corresponding estimates of these values by Gill Sans type, and values appertaining to the selected sampling units by ordinary italic type. Thus, with a quantitative character or *variate* y , the values for the selected sampling units will be denoted by y (with or without suffices as necessary), the mean of these values will be denoted by $\bar{y}$ (following the ordinary convention), the estimate of the mean of the population by $\bar{y}$, and the true mean of the population by $\bar{y}$. With a random sample we shall have $\bar{y} = \bar{y}$, but $\bar{y}$ differs from $\bar{y}$ by the sampling error. Totals for the population are indicated by capitals, summation over the sample values by S , and summation over the different strata by Σ .

In certain types of estimation we shall be concerned with the use of supplementary information, such as size of unit, which is known not only for the selected sampling units but also for the whole of the population, or in the case of two-phase sampling, for the larger number of units selected at the first phase. A variate representing quantitative supplementary information will be denoted by x .

Even when information on a variate is not available for the whole population, it may be necessary to make an estimate of the population values for some standardized value of this variate. The letter x will also be used to denote a variate of this type.

The following is a list of the principal symbols employed in this chapter :—

- b , estimated regression coefficient.
- f , working sampling fraction.
- $\mathbf{f}$, exact sampling fraction.
- g , working raising factor ($= 1/f$).
- $\mathbf{g}$, exact raising factor ($= 1/\mathbf{f}$).
- i (suffix), denotes values belonging to a particular stratum i .
- n , number of units in the sample.
- $N, \mathbf{N}$, number of units in the population, and its estimate.
- $p, \mathbf{p}$, proportion of units in the population possessing a given attribute, and its estimate.
- r , ratio y/x .
- $\bar{r}, \mathbf{\bar{r}}$, ratio $\mathbf{Y/X}$, and its estimate.
- S , summation over the units of the sample.
- S_i , summation within stratum i .
- Σ , summation over the strata.
- u , number of units in the sample possessing a given attribute.
- $U, \mathbf{U}$, number of units in the population possessing a given attribute, and its estimate.
- x , supplementary quantitative variate, such as size of unit.
- $\mathbf{X}, \mathbf{X}$, total of x for the population, and its estimate.
- y , quantitative variate under investigation.
- $\mathbf{Y}, \mathbf{Y}$, total of y for the population, and its estimate.
- $\bar{r}, \bar{x}, \bar{y}$, means of r, x, y , for the sample.
- $\bar{\mathbf{x}}, \bar{\mathbf{y}}, \bar{\mathbf{x}}, \bar{\mathbf{y}}$, means of x and y for the population, and their estimates.

6.3 General rules

There are certain fundamental rules of estimation which apply to all types of sampling. These are :—

Rule 1—The population total of a quantitative variate

To estimate totals for the population multiply all sample values by their raising factors (equal to the reciprocals of the sampling fractions) and sum the raised results.

Rule 2—Number of units in the population

To estimate the number of sampling units in the population follow Rule 1, scoring each selected sampling unit as 1.

Rule 3—The population mean of a quantitative variate

Divide the estimated total of the variate for the population by the estimated number of units in the population.

Rule 4—Proportion (or percentage) of units possessing a given qualitative character

Proceed as for a quantitative variate, scoring all units possessing the given character as 1 and all others as 0. Divide the estimated total score by the estimated number of units in the population.

Rule 5—Ratio of two quantitative variates

Estimate the totals of the two quantitative variates for the population by Rule 1 and take the ratio of these totals. (Rules 3 and 4 are special cases of Rule 5.)

In cases in which the probability of selection of all units is the same (uniform sampling fraction or, in the case of multi-stage sampling, uniform overall sampling fraction), the first four rules can be condensed into the simple general rule that means and proportions in the population are estimated by the corresponding means and proportions in the sample, and totals and numbers in the population are estimated by multiplying the corresponding totals and numbers by the common raising factor.

The above rules cover most of the methods of estimation discussed in this manual except those involving regression, which cannot easily be summarised in simple rules. They give rise to the formulæ of estimation set out in the following sections of this chapter.

6.4 Random sample

Number :

$$N = gn$$

N will be equal to N except for minor discrepancies due to the use of a working sampling fraction which does not give an integral number of sampling units. If N is known then the true sampling fraction f equals n/N and the true raising factor g equals N/n .

Mean of a quantitative variate :

$$\bar{y} = \bar{y} = \frac{1}{n} S(y) \quad (6.4.a)$$

Total of a quantitative variate :

$$Y = g S(y) = N\bar{y}$$

or more accurately, if N is known, and differs from N ,

$$Y' = g S(y) = N\bar{y} \quad (6.4.b)$$

Proportion possessing a given attribute :

$$P = \frac{u}{n}$$

Number possessing a given attribute :

$$U = gu = Np$$

or, more accurately,

$$U' = gu = Np \quad (6.4.c)$$

The same formulæ of estimation will hold for systematic samples from lists, etc.

Example 6.4.a

In a housing survey of a town a systematic sample from a list of all houses was taken with a sampling fraction of $1/50$. 627 houses out of a total of 8491 in the sample were classified as defective. What is the estimated number and percentage of defective houses in the town?

$$\text{Percentage defective} = 100 p = 100 \times \frac{627}{8491} = 7.38 \text{ per cent.}$$

$$\text{Total number defective} = U = 50 \times 627 = 31,350$$

Example 6.4.b

If the values in Table 6.4 are taken to represent measurements on a random sample of 20 objects, selected from a batch of such objects with a sampling

TABLE 6.4—SAMPLE OF 20 MEASUREMENTS

6.2	8.0	8.2	11.0
13.8	12.0	8.7	10.3
8.0	10.7	8.5	14.6
7.6	9.1	10.1	8.0
10.3	10.4	9.3	9.0

fraction of $1/25$, estimate the mean measurement of the batch, and the total of all the measurements of the batch.

$$N = 25 \times 20 = 500$$

$$S(y) = 193.8$$

$$\bar{y} = \frac{1}{20} \times 193.8 = 9.69$$

$$Y = 25 \times 193.8 = 4845$$

If the number in the batch is known to be 507, a slightly more accurate estimate of the total is

$$Y' = 507 \times 9.69 = 4913$$

6.5 Stratified sample with uniform sampling fraction

The formulæ for a random sample hold, except that if the numbers in the different strata N_i are known, and differ from N_i , the formula 6.4.b is replaced by

$$Y' = \Sigma (N_i \bar{y}_i) \quad (6.5.a)$$

and the formula 6.4.c by

$$U' = \Sigma (N_i p_i) \quad (6.5.b)$$

with corresponding slight increases in accuracy in $\bar{y}$ and p , if they are derived from these estimates by division by N .

Example 6.5

Table 6.5.a shows the wheat acreages of the stratified random sample of 1 in 20 Hertfordshire farms described in Section 3.7. Estimate the total wheat acreage of the county and the mean acreage of wheat per farm (*a*) from the data of the sample alone, (*b*) given the total number of farms in each size-group. Estimate also the number of farms growing wheat.

TABLE 6.5.a—HERTFORDSHIRE FARMS, 1939: ACREAGES OF WHEAT IN A STRATIFIED RANDOM SAMPLE OF 1 IN 20 FARMS (STRATIFIED BY ACREAGES OF CROPS AND GRASS)

Size-group .	3	4	5	6
Acres . . .	21-50	51-150	151-300	301-
No. of farms	18	26	20	13
	8 0 0 5 0 0 0 0 0 0 0 9 0 0 8 0 5 0	49 19 10 14 27 4 33 0 4 12 30 13 0 0 0 16 13 28 0 5 27 23 10 22 24 3	20 56 24 18 30 17 59 32 17 71 76 48 80 70 0 62 36 0 0 0	72 92 69 78 51 84 0 102 13 92 158 62 0
TOTAL	35	386	710	873

Size-group 1 (1-5 acres): 22 farms, no wheat.

Size-group 2 (6-20 acres): 26 farms, 7 acres of wheat on 1 farm.

The results are summarized in Table 6.5.b. The estimate of the total area of wheat in the county from the sample is

$$Y = 20 \times 2011 = 40,220 \text{ acres}$$

The mean area of wheat per farm is

$$\bar{y} = 2011/125 = 16.1 \text{ acres}$$

The number of farms growing wheat is

$$U = 20 \times 54 = 1080$$

TABLE 6.5.b—SUMMARY OF SAMPLE OF TABLE 6.5.a

Size-group, acres	No. of farms in sample	Farms with wheat in sample		Wheat acreage in sample		No. of farms in county	Total wheat acreage
		No.	Proportion	Total	Mean		
1-5	22	0	.000	0	0.0	435	0
6-20	26	1	.038	7	0.3	519	160
21-50	18	5	.278	35	1.9	357	680
51-150	26	21	.808	386	14.8	519	7,680
151-300	20	16	.800	710	35.5	400	14,200
301-	13	11	.846	873	67.2	266	17,880
ALL	125	54	.432	2,011	16.1	2,496	40,600

If the total number of farms in each size-group is known, the estimate of Y can be calculated with slightly more accuracy by using the size-group means, as shown in the last three columns. This gives an estimate Y' of the total area of wheat of 40,600 acres, and a mean area per farm $\bar{y}'$ of $40,600/2496 = 16.3$ acres. The gain in accuracy is here quite trivial, since the variation within each stratum is large relative to the mean of that stratum.

The number of farms growing wheat can be estimated similarly from the proportions in the size-groups, giving

$$U' = 0 \times 435 + 0.038 \times 519 + \dots = 1083$$

Again the gain in accuracy is trivial.

6.6 Random sample, stratified after selection

The means of, or proportions in, the different strata must be calculated separately, and formulæ 6.5.a and 6.5.b used, with division by N for estimates of $\bar{y}$ and p .

Example 6.6

Table 6.6.a shows the data, including acreages of crops and grass, for the random sample of 1 in 20 Hertfordshire farms described in Section 3.7. Estimate the total area of wheat and the number of farms growing wheat (a) directly from the sample, (b) by stratification by size, given the total numbers of farms in the size-groups of Table 6.5.b.

TABLE 6.6.a—HERTFORDSHIRE FARMS, 1939: ACREAGES OF CROPS AND GRASS (1ST COLUMN), AND OF WHEAT (2ND COLUMN), OF A RANDOM SAMPLE OF 1 IN 20 FARMS (CLASSIFIED BY DISTRICTS AFTER SELECTION)

District 1 15 farms		District 3 40 farms				District 4 24 farms		District 5 4 farms		District 6 24 farms	
188	16	370	67	40	0	11	0	4	0	8	0
60	0	26	0	28	0	6	0	312	102	87	14
192	0	369	58	221	59	543	80	8	0	6	0
48	0	212	45	31	0	822	265	11	0	44	0
44	0	153	20	6	0	654	112			4	0
79	33	287	44	34	0	3	0	335	102	614	72
14	0	28	0	316	75	158	50			192	20
465	92	14	0	116	33	4	0			10	0
197	0	4	0	4	0	68	27	District 7		24	0
163	0	17	0	409	102	55	12	10 farms		2	0
198	0	2	0	6	0	4	0			9	0
78	0	3	0	115	0	2	0			3	0
6	0	7	0	19	0	192	24	128	5	2	0
35	0	6	0	274	6	4	0	4	0	120	24
168	0	335	82	3	0	491	24	46	0	58	0
		4	0	144	0	224	28	181	20	20	0
1,935	141	1	0	3	0	280	75	17	0	30	0
		4	0	482	62	90	0	24	0	197	6
District 2		180	0	156	28	3	0	10	0	14	3
		120	11	302	71	3	0	36	0	32	6
8 farms						6	0	12	0	2	0
				4,851	763	4	0	89	0	285	29
8	0					161	80			138	0
294	29					246	60	547	25	126	0
597	107					4,034	837			2,027	174
8	0										
2	0										
200	65							GRAND TOTAL, 125 farms :—		15,114	2,301
14	0										
262	58										
1,385	259										

TABLE 6.6.b—HERTFORDSHIRE FARMS, 1939: ESTIMATION OF WHEAT ACREAGE FROM THE RANDOM SAMPLE OF 1 IN 20 FARMS (TABLE 6.6.a) STRATIFIED BY SIZE-GROUPS AFTER SELECTION

Size-group acres	No. in sample	Farms with wheat		Acreage of wheat		No. of farms in county	Total for county
		No.	Pro-portion	Total	Mean		
1-5	25	0	0	0	0	435	0
6-20	26	1	.038	3	0.1	519	50
21-50	16	1	.062	6	0.4	357	140
51-150	17	8	.471	159	9.4	519	4,880
151-300	26	20	.769	762	29.3	400	11,720
301-	15	15	1.000	1,371	91.4	266	24,310
	125	45	.360	2,301	18.4	2,496	41,100

(a) Total area of wheat = $2301 \times 20 = 46,020$ acres.

Number of farms growing wheat = $20 \times 45 = 900$.

(b) Classifying the data by size-groups (crops and grass) the numbers and totals shown in Table 6.6.b are obtained. The mean wheat acreage is then calculated for each size-group, multiplied by the total number of farms in that size-group, and the products summed, giving an estimated total wheat acreage of 41,100. Similarly, using the proportion of farms with wheat instead of the mean acreage for each size-group, the estimated number of farms growing wheat is found to be $.038 \times 519 + .062 \times 357 + \dots = 860$.

6.7 Stratified sample (variable sampling fraction)

$$N = \sum (g_i n_i)$$

$$Y = \sum (g_i S_i (y))$$

$$\bar{y} = Y/N$$

$$U = \sum (g_i u_i)$$

$$p = U/N$$

If the N_i are known, the alternative formulæ 6.5.a and 6.5.b, with division by N for $\bar{y}$ and p , are slightly more accurate.

Example 6.7

Table 6.7.a shows the Hertfordshire farm data for the stratified systematic sample with a variable sampling fraction described in Section 3.7. Estimate the total wheat acreage and the number of farms growing wheat.

TABLE 6.7.a—HERTFORDSHIRE FARMS, 1939: STRATIFIED SYSTEMATIC SAMPLE OF WHEAT ACREAGES, WITH A VARIABLE SAMPLING FRACTION (CLASSIFIED BY DISTRICTS)

Size-group : Sampling fraction : No. in sample :	1-5 Nil 0	6-20 1/200 3	21-50 1/60 6	51-150 1/20 26	151-300 1/10 40	301-500 1/5 43	501- 1/3 17
District							
1	—	—	0	0 30 6	17 18 0 28	172 0 92 56	114
2	—	0	10	0 0 40	30 16 50 55 62	50 49 121 63 72 100 186 124 105 104 160	119 107 101 160
3	—	—	0 0	25 5 10 0 0	0 0 77 0 41 42 0 24 25 61	67 22 5 58 75 51 78 94 126 86 97	195 120
4	—	0	17	28 24 8 0 5	42 0 24 54 60 75 44 6	88 65 58 94 115 98 121 80 92 18 40 120	268 260 265 260 112 155 240 168 209
5	—	—	—	0 0	22 31 32 27 0	66 142 26	—
6	—	0	0	0 0 0 0 19	38 0 0 56 17 29	0 0	72
7	—	—	—	0 0 14	0 60	16	—
TOTAL	—	0	27	214	1,163	3,292	2,925

The calculations are shown in Table 6.7.b. They follow the same lines as before, except that the sample total for each stratum must be raised separately. Using the working sampling fractions we obtain estimates of 42,765 acres of wheat and 911 farms growing wheat.

TABLE 6.7.b—SUMMARY OF SAMPLE OF TABLE 6.7.a

Size-group acres	No. in sample	No. with wheat	Total acreage	Raising factor	Raised totals		Mean acreage per farm
					No.	Acreage	
1-5	0	—	—	—	—	—	—
6-20	3	0	0	200	0	0	0
21-50	6	2	27	60	120	1,620	4.5
51-150	26	12	214	20	240	4,280	8.2
151-300	40	30	1,163	10	300	11,630	29.1
301-500	43	40	3,292	5	200	16,460	76.6
501-	17	17	2,925	3	51	8,775	172.1
	135				911	42,765	

6.8 Use of supplementary information in estimation

As already indicated, supplementary information on a quantitative character, the values of which are known for all the units of the population, can be used as the basis of stratification, or for the adjustment of an unstratified sample by stratification after selection. Alternatively, as mentioned in Section 2.8, such information can be used directly without stratification. Two methods, the *ratio method* and the *regression method*, are available. In either case only the total or mean of the supplementary variate for the whole population need be known (in addition to the values for the selected sampling units). The ratio method is simpler computationally, but the regression method is in certain circumstances more accurate.

In the ratio method, the ratio of Y/X in the population is estimated from the sample, the estimated ratio being multiplied by the total X of x for the population to give the estimated total Y of y for the population. The method of estimation must be such that bias is avoided. As already explained in Section 2.6, the appropriate estimate of the ratio for a random sample is $S(y)/S(x)$ or $\bar{y}/\bar{x}$. More generally, Rule 5 of Section 6.3 will give an unbiased estimate, though in certain cases separate values of the ratio may be estimated for the different strata as described in Sections 6.10 and 6.11.

In the regression method the average change of y for unit change of x (known as the *regression coefficient*) is estimated, and this coefficient is used to adjust the sample results for any discrepancy between the mean size of unit in the sample and in the population.

The contrast between the ratio and regression methods is illustrated in Fig. 6.8. The data plotted are those of Table 6.12. The dots represent the x and y values of the sample points, the sample mean $(\bar{x}, \bar{y})$ being M . Q' represents the known population mean $\bar{x}$ of the supplementary variate, which differs from $\bar{x}$ by QQ' . The line OMD through the origin and the mean represents the ratio $\bar{y}/\bar{x}$ given by the sample, and the ordinate $P_1 Q'$ of the point P_1 on this line, equal to $(\bar{y}/\bar{x}) \bar{x}$, gives the adjusted estimate of the population mean by the ratio method. The regression line AMB also passes through the mean, and has a slope b equal to the regression coefficient.

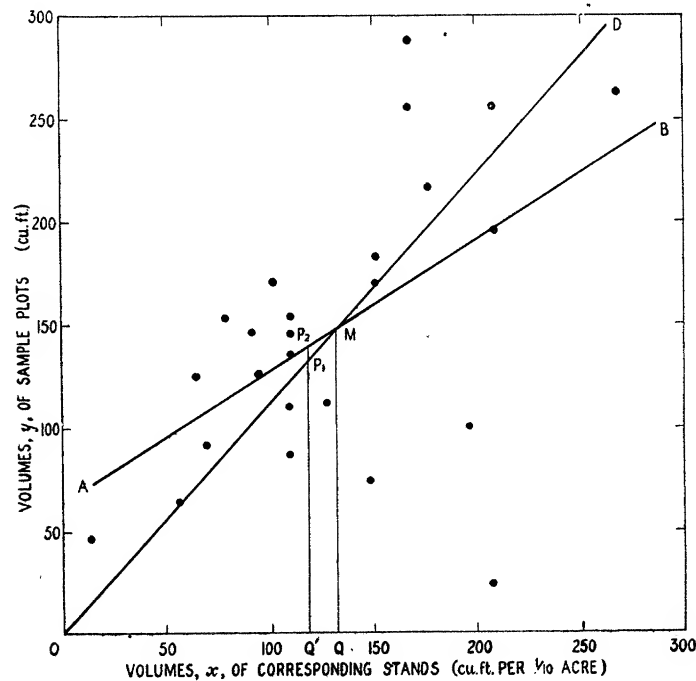


FIG. 6.8—USE OF SUPPLEMENTARY INFORMATION: RATIO AND REGRESSION METHODS (DATA OF TABLE 6.12)

This line has the property that the sum of the squares of the vertical distances of the sample points from it is minimum. The adjusted estimate by the regression method is given by the ordinate $P_2 Q'$ of the point P_2 , and equals $\bar{y} + b(\bar{x} - \bar{x})$.

The regression method therefore differs from the ratio method in that in the former the straight line which best fits the sample values is taken, whereas in the latter the line through the origin is taken. When the supplementary variate x represents size of unit, the true regression line generally passes

through the origin, though curvature of this line may result in the best-fitting straight regression line not passing through, or even very close to, the origin. Nevertheless, in most census work in which x represents size, or some variate closely correlated with size, the greater simplicity of the ratio method outweighs any small gain in accuracy resulting from the use of regression.

It may be noted that in large samples the regression line can be plotted by grouping the data according to the x values and plotting the means of y for the different groups.

The formulæ for regression have been included in this book, not because it is expected they will be very commonly used in census work, but because the regression method represents an important part of sampling procedure, without which no account of sampling methods would be complete, and because the calculation of the sampling errors to which a balanced sample is subject can only be made by use of the regression concept.

If the population mean of x is estimated from observations at the first phase of two-phase sampling, the observations for y being obtained at the second phase, the same formulæ of estimation hold, the estimate $\bar{x}_1$ being substituted for $\bar{x}$.^{*} If, however, the sampling is single-phase, and the estimate $\bar{x}$ from the sample is substituted for $\bar{x}$, the formula $y = \bar{y}$ appropriate to a random sample without supplementary information is obtained. In other words, there is no gain in accuracy in the population estimate unless x is known or alternatively is estimated from a larger sample than is available for y .

In addition to their use in the adjustment of the population estimate of the mean or total of y when the mean or total of x is known for the population, or is determined from the first phase of two-phase sampling, ratios and regressions are of use for the purpose of obtaining estimates of the means of y for some standardized value of x . Hence comparisons of different parts of the population can be made, freed from the effects of variation in the average values of x . In the case of ratios this is equivalent to comparing the values of ratios themselves: thus in an agricultural survey we may consider such quantities as number of sheep per 100 acres instead of number of sheep per farm. Regression enables similar standardization to be made in cases in which the ratio method is inappropriate; in a nutrition survey, for example, it may well be found that the amount of malnutrition varies with size of family, but the relation will not be proportionate. In large-scale surveys, however, standardization of this type can equally well be made by using the size-group means, thus avoiding the trouble of calculating regression coefficients.

The formulæ in the following sections are given for a quantitative variate. Formulæ for the proportion of units possessing a given attribute can be derived by scoring each unit as 1 if it has the attribute and zero otherwise. If the supplementary variate x represents size, the proportion of the attribute per unit size may be required, in which case a unit of size x will be scored x if it has the attribute, and zero otherwise.

^{*} Note that the observations on x for the units for which y is also observed are part of the first-phase information, and must be included when calculating this estimate.

Example 6.8

The data of Table 6.8 are extracted from the *Report of the National Farm Survey of England and Wales* (Ministry of Agriculture and Fisheries, 1946, G). They give the rents of holdings per acre of crops and grass, classified by size-groups, in Berkshire and Cornwall. Calculate rents per acre standardized for size of holding, in the proportions in which the different size-groups occur in the whole country.

TABLE 6.8—RENTS PER ACRE FOR BERKSHIRE AND CORNWALL

Size-group acres	Rent per acre (shillings)		Proportionate areas of different size-groups in whole country
	Berkshire	Cornwall	
5-25	53	55	1
25-100	32	31	6
100-300	24	22	10
300-700	20	18	4
700-	17	14	1
Overall	23	28	22

The proportionate areas for the whole country are shown in the last column. The standardized rent for Berkshire, using these areas as weights, is $(1 \times 53 + 6 \times 32 + \dots)/22 = 26$ shillings per acre, and that for Cornwall is 25 shillings per acre.

Inspection of the table shows that although the overall rent per acre is considerably less for Berkshire than for Cornwall, there is little difference between the two counties for the different size-groups, the rents for Berkshire being in general somewhat greater than those for Cornwall. This is brought out, in a single contrast, by the standardized rents. The lower overall rent per acre for Berkshire is in a certain sense accounted for by the greater proportion of large farms in that county.

This example illustrates both the use and the danger of standardization of this type. The standardized rents eliminate the effects of differences in average size on the average rent, which in so far as they are due to greater concentration of buildings on the smaller holdings, greater demand for smaller holdings, etc., do not represent differences in value of the land. It would be incorrect, however, to assume that were the size-distributions of farms in the two counties the same the overall rents per acre would be the same. Part of the difference

is due to the tendency of poorer land to be farmed in larger units, and such land would not command the full increase in rent which is apparently attracted by smaller farms if it were divided into smaller units.

6.9 Ratio method : random sample

$$\bar{y} = \frac{S(y)}{S(x)} \bar{x} = \frac{\bar{y}}{\bar{x}} \bar{x} \quad (6.9)$$

$$Y = \frac{S(y)}{S(x)} X$$

$$\bar{r} = \frac{S(y)}{S(x)}$$

The formula for $\bar{y}$ may be used for obtaining the "standardized" value y_0 of y for a standard value x_0 of x , or for estimation in two-phase sampling using an estimate $\bar{x}_1$ obtained from first-phase information.

Example 6.9.a

Estimate the total area of wheat from the data of Example 6.6 by the ratio method, given that the total area of crops and grass in the county is 273,074 acres.

Crops and grass in sample = $S(x) = 15,114$ acres

Wheat in sample = $S(y) = 2301$ acres

Estimate of wheat acreage in county = $\frac{2301}{15,114} \times 273,074$
 $= 0.15224 \times 273,074$
 $= 41,570$ acres.

Example 6.9.b

Table 6.9 gives the numbers of persons belonging to 43 kraals which form a random sample of the 325 kraals in the Mondora Reserve in Southern Rhodesia, and also the numbers of persons absent from these kraals. (The

TABLE 6.9—DATA FROM A RANDOM SAMPLE OF 43 KRAALS: TOTAL NUMBER OF PERSONS (INCLUDING ABSENTEES), x , AND NUMBER OF ABSENTEES, y

x	y	x	y	x	y	x	y
95	18	89	7	75	12	159	36
79	14	57	9	69	16	54	26
30	6	132	26	63	9	69	27
45	3	47	7	83	14	61	2
28	5	43	17	124	25	164	69
142	15	116	24	31	3	132	41
125	18	65	16	96	45	82	10
81	9	103	18	42	25	33	8
43	12	52	16	85	35	86	22
53	4	67	27	91	28	51	19
148	31	64	12	73	13		
						3,427	799

data form part of the results of a sample census of the Hartley District, and have kindly been made available by Dr. J. R. H. Shaul.) Estimate the percentage of persons absent from the reserve, and the numbers of persons belonging to the reserve and absent from the reserve.

$$\text{Percentage absent} = 100 \bar{r} = \frac{799}{3427} \times 100 = 23.3 \text{ per cent.}$$

$$\text{Number belonging to reserve} = X = \frac{325}{43} \times 3427 = 25,902.$$

$$\text{Number absent from reserve} = Y = \frac{325}{43} \times 799 = 6039.$$

$$\text{Number present in reserve} = 25,902 - 6039 = 19,863.$$

This example, though superficially similar to Example 6.4.a, is structurally different, in that the sampling units consist of kraals and not individuals. This affects the estimation of the sampling error (see Section 7.8).

6.10 Ratio method : stratified sample with uniform sampling fraction

(a) When the ratio is assumed to be the same for all strata :
The formulæ for a random sample hold.

(b) When the ratio is permitted to assume different values for the different strata :

Treat each stratum separately, using the formulæ for a random sample, and build up the population estimates by summation of the estimates for the separate strata, with division by N or N for the population means.

This gives

$$Y = \Sigma \left(\frac{S_i(y)}{S_i(x)} X_i \right) \quad (6.10)$$

etc.

The choice between method (a) and method (b) depends on :

- (1) Numbers in the different strata—method (b) can only be used if the numbers of units from the individual strata are sufficiently large to give reasonably accurate determinations of the values of ratio for the separate strata : if the numbers are small and there is correlation between r and x , the method will be biased. (This objection does not hold if selection with probabilities proportional to x is used—see Section 3.10.)
- (2) The degree of variation in the ratio between the different strata—the greater this variation the greater will be the gain in accuracy by the use of method (b) ; if the variation is small, method (a) may be the more accurate.

- (3) Computational convenience—method (a) is simpler, since only one value of the ratio is involved.

Method (b) can also be used for a random sample stratified after selection, provided the population totals of x for each stratum are known.

Example 6.10

Estimate the total area of wheat from the data of Example 6.6 by the ratio method, stratifying the data by districts, and using different values of the ratio for the different districts.

TABLE 6.10—HERTFORDSHIRE FARMS, 1939: ESTIMATION OF WHEAT ACREAGES FROM THE RANDOM SAMPLE OF 1 IN 20 FARMS BY THE RATIO METHOD AFTER STRATIFICATION INTO DISTRICTS

District No.	Sample				District crops and grass X_i	Estimated district wheat $r_i X_i$
	No.	Wheat $S_i(y)$	Crops and grass $S_i(x)$	Ratio r_i		
1	15	141	1,935	·0729	22,932	1,670
2	8	259	1,385	·1870	43,591	8,150
3	40	763	4,851	·1573	57,263	9,010
4	24	837	4,034	·2075	73,946	15,340
5 & 7	14	127	882	·1440	40,905	5,890
6	24	174	2,027	·0858	34,437	2,950
	125	2,301	15,114		273,074	43,010

The computations are shown in Table 6.10. The neighbouring districts of St. Albans (5) and Watford (7) (each of which contains rather a small number of farms) have been combined.

6.11 Ratio method : stratified sample with variable sampling fraction

- (a) Ratio the same for all strata :

$$\bar{r} = \frac{\sum \{g_i S_i(y)\}}{\sum \{g_i S_i(x)\}}$$

$$\bar{y} = \bar{r} \bar{x}$$

$$Y = \bar{r} X$$

161
PROPERTY OF
CARNEGIE INSTITUTE OF TECHNOLOGY
LIBRARY

(b) Ratio different for different strata :

Proceed in the same manner as for a fixed sampling fraction. The sampling fraction does not enter into the calculations.

6.12 Regression method : random sample

The equation of the regression line is

$$y_l = \bar{y} + b(x - \bar{x})$$

where

$$b = \frac{S(y - \bar{y})(x - \bar{x})}{S(x - \bar{x})^2} \quad (6.12.a)$$

Hence

$$\bar{y} = \bar{y} + b(\bar{x} - \bar{x})$$

$$Y = N \bar{y} \quad (6.12.b)$$

If N is not known exactly, it must be estimated from the sample.

Note that if b is put equal to $S(y)/S(x)$ formula 6.9 is obtained, and if put equal to 0 formula 6.4.a is obtained. All values of b will give unbiased estimates, and consequently any value b_0 which appears appropriate to the data under analysis may be used. Thus, taking $b_0 = 1$ is equivalent to the use of the differences $y - x$. The regression method furnishes the value of b which gives the most accurate estimate of $\bar{y}$, using a formula of type 6.12.b, at the cost of some additional computational labour.

Regressions may be used for standardization and in two-phase sampling in the same way as ratios.

Example 6.12.a

Obtain an estimate of the total area of wheat from the data of Example 6.6, using the regression method.

We have

$$\begin{array}{llll} \bar{x} = 120.912 & \bar{y} = 18.408 & N = 2496 & \bar{x} = 109.405 \\ S(x^2) = 5,061,734 & S(xy) = 902,958 & & S(y^2) = 207,261 \\ \bar{x} S(x) = 1,827,464 & \bar{y} S(x) = \bar{x} S(y) = 278,219 & & \bar{y} S(y) = 42,357 \\ S(x - \bar{x})^2 = 3,234,270 & S(x - \bar{x})(y - \bar{y}) = 624,739 & & S(y - \bar{y})^2 = 164,904 \end{array}$$

The method of calculation of the sums of squares and products is explained in Section 7.1. The sum of squares of y will be required in the calculation of the sampling error.

We then have

$$b = \frac{624,739}{3,234,270} = 0.19316$$

$$\bar{y} = 18.408 + 0.19316(109.405 - 120.912) = 16.185$$

$$Y = 2496 \times 16.185 = 40,400 \text{ acres}$$

Example 6.12.b

Table 6.12 gives the measured volumes of timber on 25 systematically located plots of 1/10 acre, and eye estimates of the volumes per 1/10 acre in

TABLE 6.12—MEASURED VOLUMES, y , ON 25 SAMPLE PLOTS, AND EYE ESTIMATES, x , OF CORRESPONDING STANDS (CU. FT. PER 1/10 ACRE)

y	x	y	x	y	x	y	x
170	102	195	208	153	79	169	152
47	14	255	208	216	177	182	152
64	57	135	110	125	65	74	148
91	70	146	110	100	196	24	207
126	95	154	110	287	167	255	167
146	92	110	110	261	268	3,684	3,302
87	110	112	128			147.36	132.08

the stands in which they occurred. If more than one sample plot occurs in a stand this is indicated by a bracket, but the observations have been treated as independent in the subsequent computations. The data, which refer to conifer stands of uniform age and over 20 years of age in two counties, were obtained in the course of the 1938-9 Census of Woodlands. They are plotted in Figure 6.8. The total area of conifer stands over 20 years of age in these two counties was 5124 acres, and the total volume of timber, from eye estimates of all these stands, was 6,110,000 cu. ft., *i.e.* 1192 cu. ft. per acre. Obtain unbiased estimates of the total volume of this class of timber from the above data.

The mean of the measured volumes on the sample plots provides an unbiased estimate of the volume per acre, and the estimate of the total volume, based on the measurements of the sample plots only, is consequently

$$147.36 \times 10 \times 5124 = 7,551,000 \text{ cu. ft.}$$

The ratio method gives

$$1192 \times 5124 \times 1473.6 / 1320.8 = 6,814,000 \text{ cu. ft.}$$

Elimination of possible bias in the eye estimates by taking the difference between the measured volumes and eye estimates on the sample plots (equivalent to the use of the regression method with an arbitrary coefficient $b_0 = 1$) gives

$$5124 (1473.6 - 1320.8 + 1192) = 6,891,000 \text{ cu. ft.}$$

The regression method proper requires the calculation of the regression coefficient. We find

$$S(y - \bar{y})^2 = 115,266 \quad S(y - \bar{y})(x - \bar{x}) = 52,069 \quad S(x - \bar{x})^2 = 82,296$$

$$b = 52,069 / 82,296 = 0.63270$$

$$Y = 5124 \{ 1473.6 + 0.63270 (1192 - 1320.8) \} = 7,133,000 \text{ cu. ft.}$$

The relative accuracy of these various methods of estimation is discussed in Example 7.12.b.

The above data are of course only a small part of the full data for the survey. Examination of the whole of the data for the above two counties gave a value of b of 0.55. The bias in the eye estimates, which are too low, though not very large, is apparent in the above data. The average bias over the whole survey was decidedly larger, and misleading results would have been obtained by using the eye estimates without correction for bias from properly measured and randomly located sample plots.

There is of course the possibility—if the location of the sample plots has not been objectively carried out, or if the measurements have been carelessly made, *e.g.* by the inclusion of trees whose centres do not lie within the demarcated sample area—that the sample plots will themselves be biased. The sample plots used in this survey were somewhat small, and the use of larger plots, particularly in the case of hardwoods, possibly with second-stage sampling of trees for measurement, would have reduced the risk of bias of this nature. The surveyors were well trained, however, and thoroughly appreciated the need for objectivity, and on examination it appeared that serious bias from this cause could be ruled out. The results of a later survey of England and Wales confirmed the correctness of the earlier survey.

6.13 Regression method : stratified sample with uniform sampling fraction

(a) When the regression coefficient is assumed to be the same for all strata :

The formulæ for a random sample hold, except that formula 6.12.a is replaced by

$$b = \frac{\sum \{S_i (y - \bar{y}_i) (x - \bar{x}_i)\}}{\sum \{S_i (x - \bar{x}_i)^2\}}$$

(b) When the regression is permitted to assume different values in the different strata :

Proceed as in the ratio method.

6.14 Regression method : stratified sample with variable sampling fraction

(a) Regression the same for all strata :

$$\bar{y} = \bar{y}_w + b (\bar{x} - \bar{x}_w)$$

where

$$b = \frac{\sum \{ \lambda_i S_i (y - \bar{y}_i) (x - \bar{x}_i) \}}{\sum \{ \lambda_i S_i (x - \bar{x}_i)^2 \}}$$

the λ_i being numerical weighting coefficients, and

$$\bar{y}_w = \frac{\sum \{ g_i S_i (y) \}}{\sum (g_i n_i)}$$

with a similar expression for $\bar{x}_w$. $\bar{y}_w$ and $\bar{x}_w$ are the estimates of $\bar{y}$ and $\bar{x}$ that would be obtained from the sample if there were no supplementary information on x (see Section 6.7).

If the regressions within strata are truly linear, with identical values of the regression coefficient, then the most accurate estimate of b will be obtained if the λ_i are taken inversely proportional to the residual within-strata variances of y about the regression lines. If the regression coefficients are different for the different strata, then the component of error due to the assumption of equality of regression coefficients will be minimized by taking λ_i proportional to g_i^2 . Any set of λ_i will give a virtually unbiased estimate of $\bar{y}$, and detailed investigation of the theoretically best values to adopt is seldom worth while. For most work λ_i may be taken as unity if all the strata contain material of similar variability, *i.e.* if the variable sampling fraction arises from extraneous causes not connected with the variability of the material, and equal to g_i if the sampling fractions have been chosen so as to minimize the sampling error. Under certain conditions $\lambda_i = g_i^2$ would be best in this case, but under other conditions this would give excessive weight to the strata with small sampling fractions.

(b) Regression different for different strata :

Proceed as in the ratio method.

6.15 Use of regression to calibrate eye estimates

It sometimes happens that eye estimates or similar subjective measurements, x , can be made on a properly selected and unbiased sample of the population, but that the objective measurements y , which are required to calibrate these estimates, can only be carried out on a non-random sub-sample of the original sample. The eye estimates cannot then be used as supplementary information in the manner of Example 6.12.b, since any bias in the sub-sample used for the objective measurements would be reflected to a greater or less extent, depending on the value of b , in the population estimate derived from the regression.

In this case the regression of x on y , instead of y on x , must be calculated, and the equation of estimation must be replaced by

$$\bar{y} = \bar{y} + \frac{1}{b'}(\bar{x}_1 - \bar{x}),$$

when b' is the regression coefficient of x on y , $\bar{y}$ and $\bar{x}$ are the means for the sub-sample, and $\bar{x}_1$ is the mean of the eye estimates for the original sample.

This procedure is subject to certain limitations. Firstly, the sub-sample, though non-random for the whole of the population, must be effectively random for units having any given value of y . If, for example, there is a tendency to select units which, for a given value of y , have high values of x , serious bias may result. Thus, in a crop-estimation scheme, if eye estimates are made on a random sample of fields, and if reliance is placed on returns

by farmers of the actual yields of some of these fields, any tendency on the part of the farmers to return only the yields of fields which have turned out better than their appearance would indicate will lead to an overestimate of the yield. On the other hand, the omission of a greater proportion of the low-yielding than of the high-yielding fields from the sub-sample will not bias the results, provided this omission is conditioned only by the final yield and not by the previous appearance or the value of the eye estimate.

Secondly, for accuracy in the final estimate, the eye estimates must be reasonably accurate in the sense that variation about the regression line must be small, and the line itself must have an adequate slope. If the regression line is curved, this curvature can only be allowed for in the estimation formula if the variation about the regression line is negligible. Otherwise bias will be introduced. The use of the best fitting linear regression line, however, will avoid this source of bias.

Example 6.15

In order to test the accuracy of eye estimation as a method of estimating the yields of cereal crops shortly before harvest, a trial survey of the wheat crop of Hertfordshire was undertaken in 1940. Two observers were employed, one of whom visited 47 farms, observing 110 fields, and the other 16 farms, observing 37 fields. The whole set of farms constituted a systematic sample of 1 in 12 farms, excluding those growing less than 5 acres of wheat in 1939, a random sub-sample of fields being taken on the larger farms. The actual yields, as determined by the farmers, were subsequently obtained for as many of the observed fields as possible, and these were used to calibrate the eye estimates. The relation between the eye estimates, x , and the actual yields per acre, y , for the first observer are shown in Fig. 6.15 for the 37 fields for which yields were obtained. Obtain an estimate of the mean yield per acre for the part of the county covered by this observer.

The regression coefficient, b' , of x on y , calculated from the unweighted values of x and y for the 37 fields, is 0.6926, the regression equation being

$$x_i = 30.00 + 0.6926(y - 28.78)$$

This is shown by the full line in the figure. The dotted line represents the line that would be obtained if there were no errors in the eye estimates. It will be seen that there is a tendency to underestimate high yields and overestimate low yields. The other observer and the farmers gave very similar results.

The mean of the eye estimates $\bar{x}_1$ for the whole of the first observer's sample, and that for the eye estimates $\bar{x}$ and the yields $\bar{y}$ of the fields for which yields were available, weighted according to the acreages of the fields, with an additional raising factor if all the fields on the farm are not sampled, are, in bushels per acre,

$$\bar{x}_1 = 30.12 \quad \bar{x} = 30.13 \quad \bar{y} = 28.95$$

Hence, since $1/0.6926 = 1.444$, the final estimate of the yield per acre is

$$\bar{y} = 28.95 + 1.444(30.12 - 30.13) = 28.94$$

The adjustment is here negligible, since $\bar{x}_1$ and $\bar{x}$ are almost identical.

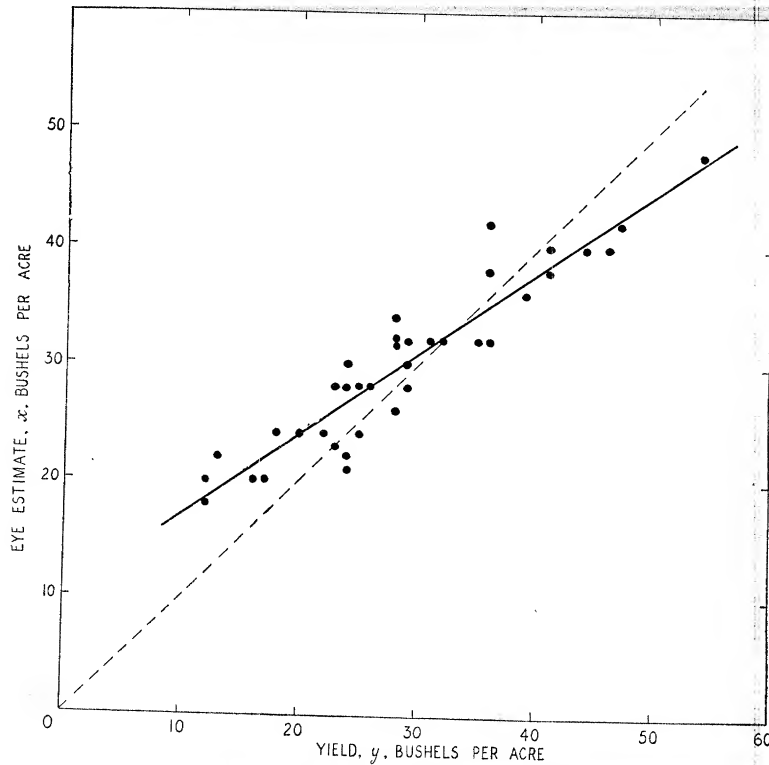


FIG. 6.15—RELATION BETWEEN EYE ESTIMATES OF THE YIELDS AND THE ACTUAL YIELDS OF 37 FIELDS OF WHEAT IN HERTFORDSHIRE

6.16 Sampling with probabilities proportional to size of unit

(a) Size, x , of all units of the population known, or $\mathbf{X}$ known:

In this case x acts as a supplementary variate, and the ratio method will in general be appropriate. Since the probability of selection is proportional to x , raising factors proportional to $1/x$ must be introduced into the formulæ already given. This leads to the formulæ

$$\bar{r} = \frac{1}{n} S \left(\frac{y}{x} \right) = \bar{r}$$

$$Y = \bar{r} \mathbf{X}$$

$$\bar{y} = \bar{r} \bar{x}$$

In other words the unbiased estimate of the population value of the ratio is given by the arithmetic mean of the ratios from the selected sampling units.

(b) Total size X of the population not known :

In this case X , as well as Y and $\bar{r}$, have to be estimated from the sample. Selection has to be made by some such process as randomly or systematically locating points on a map, and points not falling in the units under consideration must be taken into account. If n_0 is the total number of sampling points, and A is the total area covered by the sampling grid, we have

$$\begin{aligned}\bar{r} &= \bar{r} \\ X &= A n/n_0 \\ Y &= \bar{r} X = \bar{r} A n/n_0\end{aligned}$$

Alternatively, if A is not known exactly the density d of points per unit area may be used. We have

$$\begin{aligned}A &= n_0/d \\ X &= n/d\end{aligned}$$

If the sampling is two-phase, with n'_0 points (density d') at the first phase, of which n' fall in the units under consideration, n , n_0 and d must be replaced by n' , n'_0 and d' in the above formulæ.

Example 6.16.a

In a survey to estimate the area and yield of a crop, systematically located points at a density of one per 4 square miles are taken, and the yields per acre of the fields in which the points fall and which carry the crop are determined by the harvesting of small areas. 8317 points in all are obtained, of which 529 fall in fields carrying the crop. The arithmetic mean of the yields per acre of the selected fields is 15.7 cwt. per acre. Estimate the total area and yield of the crop.

A density of 1 per 4 square miles is equal to $1/2560$ per acre. Hence

$$\text{area} = X = 529 \times 2560 = 1,354,000 \text{ acres.}$$

$$\text{yield} = Y = 15.7 \times 1,354,000 \text{ cwt.} = 1,063,000 \text{ tons}$$

Example 6.16.b

If, in addition to the yield data of Example 6.16.a, a further 24,938 points were surveyed for type of crop only, giving an overall density, with the 8317 points of the yield survey, of one point per square mile, and 1673 of the fields so located were found to carry the crop in question (in addition to the 529 fields above), obtain revised estimates of the total area and yield of the crop.

This is an example of two-phase sampling. The two sets of points together constitute the first phase, for area of crop, and the 529 points for which yield samples were taken constitute the second phase.

We therefore have

$$n_0' = 24,938 + 8317 = 33,255$$

and

$$n' = 1673 + 529 = 2202$$

The density at the first phase is $1/640$ per acre. Consequently

$$\text{area} = X = 2202 \times 640 = 1,409,000 \text{ acres}$$

$$\text{yield} = Y = 15.7 \times 1,409,000 \text{ cwt.} = 1,106,000 \text{ tons}$$

6.17 Sampling from within strata with probabilities proportional to size of unit

In this case the sizes of all units will be known. As pointed out in Section 3.10, if more than one unit is selected from some or all of the strata, the same unit not being selected twice, the probabilities will not in fact be exactly proportional to size, and slight bias will be introduced. The ratios from the selected units are meaned separately for each stratum, giving equations of estimation

$$\bar{r}_i = \frac{1}{n_i} S_i \left(\frac{y}{x} \right) = \bar{r}_i$$

$$Y = \sum \{ \bar{r}_i X_i \}$$

$$\bar{r} = Y/X$$

Example 6.17

Estimate the acreage of wheat and the number of farms growing wheat in Hertfordshire from the sample of parishes described in Section 3.11.

TABLE 6.17.a—HERTFORDSHIRE WHEAT: SAMPLE OF 17 "COMBINED" PARISHES SELECTED FROM WITHIN DISTRICTS WITH PROBABILITY PROPORTIONAL TO SIZE

	1				2				3			
District . . .												
Wheat . . .	164				766	701	503		311	228	249	686
Crops and grass	3,350				3,040	3,440	2,040		2,370	3,330	2,290	2,930
Ratio . . .	.049				.252	.204	.247		.131	.068	.109	.234
	4					5			6		7	
District . . .												
Wheat . . .	558	775	495	565	862	818			225	738		290
Crops and grass	2,300	4,430	2,890	2,420	4,160	3,470			2,520	3,740		3,060
Ratio . . .	.243	.175	.171	.233	.207	.236			.089	.197		.095

TABLE 6.17.b—ESTIMATION OF WHEAT ACREAGE FROM THE DATA OF TABLE 6.17.a

District	Mean ratio wheat/crops and grass	Acreage of crops and grass in district	Estimated acreage of wheat
1	.049	22,932	1,120
2	.234	43,591	10,200
3	.136	57,263	7,790
4	.206	73,946	15,230
5	.236	24,964	5,890
6	.143	34,437	4,920
7	.095	15,941	1,510
TOTAL		273,074	46,660

The data from the sampled parishes are shown in Table 6.17.a, and the further computations for wheat acreage in Table 6.17.b.

The computations for number of farms growing wheat follow exactly the same lines, using the ratio of number of farms growing wheat to acreage of crops and grass for each parish. These computations are left as an exercise to the reader.

The use of the ratio of number of farms growing wheat to the total number of farms in the district would be equally admissible if the selection of parishes had been made with constant probability, but when the probability of selection is taken proportional to size of unit, the ratio to size, however defined, must be taken for all variates.

6.18 Multi-stage sampling, no supplementary information

In multi-stage sampling the process of estimation can be carried out stage by stage, using the appropriate methods of estimation at each stage. It is often more convenient, however, to combine all stages in a single process of estimation.

Thus the combined or *overall* raising factor g for any sub-unit in two-stage sampling is given by the product of the first-stage raising factor g' of the main unit in which it occurs and the second-stage raising factor g'' of the particular sub-unit, *i.e.*

$$g = g' g''$$

Hence the general formula for Y , when there is no supplementary information, is

$$Y = S(gy)$$

where the summation is taken over all units, with similar formulæ for N , etc.

If the combined raising factors for a group of units are equal then the computations will be simplified by summing these units before multiplication. In particular, if all the combined raising factors are equal, the sample can be treated for purposes of estimation as if it were an ordinary random or stratified random sample with uniform sampling fraction.

6.19 Multi-stage sampling with supplementary information

(a) Ratio method, ratio the same for whole population :

$$\bar{y} = \frac{S(gy)}{S(gx)} \bar{x}$$

etc., where $g = g' g''$.

(b) Ratio method, ratio different for different parts of the population :

Many variants are possible. All can be resolved by proceeding stage by stage by the methods already outlined. The danger of introducing bias if the number of units on which the ratios are based is small must be recognized.

(c) Regression method :

Regressions will usually be employed at the first stage of the sampling, in which case the regression coefficient or coefficients will be calculated in the manner appropriate to the type of sampling adopted at this stage, using the values of the totals of x and y for each main unit estimated from the second-stage sampling.

If regression is used at the second stage the procedure for stratified samples can be used, treating the selected first-stage units as if they were strata.

(d) Sampling with probability proportional to size :

An important case is that in which the first-stage units are sampled from within strata with probability proportional to size, and the second-stage sampling fractions are chosen so as to give a uniform overall sampling fraction. In this case the use of the mean ratios $\bar{r}_i'$ at the first stage of the estimation process (Section 6.17) will be found to be equivalent to the direct estimation from the second-stage units by means of the overall raising factors, *i.e.* $\bar{y} = S(gy)$.

Example 6.19

From the data of Table 6.19.a, obtained in the course of the Survey of Fertilizer Practice, estimate the average dressing of nitrogenous fertilizer on sugar beet in Norfolk.

The two-stage sampling procedure of this survey has been described in Section 4.23. Information is lacking from a few farms, mainly owing to changes in tenancy. Since this affects the small farms to a greater extent than the large farms the adjusted sampling fractions shown in the table have been used. These are equal to the number of farms on which information is available divided by the total number of farms in the size-group. The

second-stage sampling fractions are given by the reciprocals of the number of fields, since one field is selected on each farm. The combined raising factor for the sampled field of the first farm in Table 6.19.a is therefore 105, that for the sampled field of the third farm is $105 \times 3 = 315$, etc.

TABLE 6.19.a—SURVEY OF FERTILIZER PRACTICE: DATA ON THE APPLICATION OF NITROGENOUS MANURES TO SUGAR BEET ON OLD ARABLE LAND IN NORFOLK

No. of fields	Acreage		Cwt. N per acre	No. of fields	Acreage		Cwt. N per acre	No. of fields	Acreage		Cwt. N per acre
	Total Sample				Total Sample				Total Sample		
Small farms				Medium farms				Medium farms (<i>contd.</i>)			
1	2	2	.68	2	13	10	.42	1	12	12	.16
1	6	6	.63	4	40	5	.63	1	6	6	.30
3	5	2	.55	1	8	8	.15	2	14	7	.42
2	4	3	.42	3	14	4	.42	3	21	6	.14
1	5	5	.36	2	51	11	.63	1	8	8	.42
1	3	3	.21	2	16	10	.90	2	10	2	.49
1	4	4	.21	3	19	7	.21	1	4	4	.30
1	2	2	.42	3	31	10	.84	1	4	4	.30
2	6	2	.42	3	39	4	.63	3	14	7	.45
2	8	4	.90	2	19	13	.52	6	42	10	.36
1	2	2	.52	1	9	9	.52	1	6	6	.21
2	4	3	.15	1	20	20	.42	1	4	4	.30
1	6	6	.70	5	26	7	.30	2	19	8	.42
2	5	4	.36	2	8	6	.54	Large farms			
2	6	4	.56	4	20	8	.21	1	8	8	0
1	2	2	0	1	4	4	.42	1	48	48	0
1	4	4	.21	1	20	20	.72	3	19	4	0
1	5	5	.48	2	7	4	.36	3	56	24	.68
2	3	2	.10	4	32	11	.63	2	22	5	.36
1	7	7	.32	2	16	8	.82	4	30	5	.42
1	2	2	.80	2	6	3	.63	3	20	5	.90
1	4	4	.30	2	20	10	.56	6	126	29	.75
				1	7	7	.57	6	28	6	.63

Number of farms without sugar beet on old arable land: small (6–50 acres), 8; medium (51–300 acres), 11; large (301– acres), 5.

Sampling fractions (adjusted for absence of information): small, $1/105$; medium, $1/59$; large, $1/30$.

The average dressing of nitrogen must be obtained by calculating the raised total of the amount of nitrogen applied $S(gy)$ and the raised total of the acreage sampled $S(gx)$. The amount of nitrogen applied to a field is given by the product of the acreage and the rate per acre. The three size-groups are best kept separate in the computation. For the first size-group, therefore, applying the second-stage raising factors, we have

$$\begin{aligned}
 S(g''y) &= S(g''rx) = 1 \times 2 \times 0.68 + 1 \times 6 \times 0.63 + 3 \times 2 \times 0.55 + \dots \\
 &= 1.36 + 3.78 + 3.30 + \dots \\
 S(g''x) &= 1 \times 2 + 1 \times 6 + 3 \times 2 + \dots \\
 &= 2 + 6 + 6 + \dots
 \end{aligned}$$

This gives the results shown in Table 6.19.b. Applying the first-stage raising factors to the total nitrogen and total acreage, we obtain the average dressing of nitrogen per acre :

$$\bar{r} = \frac{46.13 \times 105 + \dots}{104 \times 105 + \dots} = \frac{28,512.04}{58,229} = .490 \text{ cwt. per acre}$$

TABLE 6.19.b—ESTIMATION OF AVERAGE DRESSING FROM RAISED RESULTS

Size-group	Total nitrogen $S(g'y)$	Total acreage $S(g'x)$	Nitrogen per acre $S(g'y)/S(g'x)$	First-stage raising factor g'
Small . .	46.13	104	.444	105
Medium . .	285.41	601	.475	59
Large . .	227.64	395	.576	30

The data presented comprise only a small part of the information collected, and the above method of estimation therefore demands a good deal of computation. For certain purposes comparative figures may be obtained from the straight averages of the dressings per acre on the sampled fields.

TABLE 6.19.c—ESTIMATION OF AVERAGE DRESSING FROM UNWEIGHTED MEANS

Size-group	No. of farms	Nitrogen per acre	
		Sum	Mean
Small . .	22	9.30	.423
Medium . .	36	16.32	.453
Large . .	9	3.74	.416
All . .	67	29.36	.438

These averages are given in Table 6.19.c. The first-stage raising factors have not been used in this calculation, since the larger farms have more and larger fields in sugar beet, so that inequality in the omitted second-stage raising factors more than compensates for the difference in the first-stage raising factors.

It will be noted that the mean dressings are less than those previously obtained for all size-groups, indicating the possibility that farms with little sugar beet, which are overweighted in the straight averages, are using less nitrogen per acre than those with a large amount of sugar beet. The data are too variable, however, to determine with certainty from this sample alone whether this is really a bias or is due to random sampling errors. The large difference for the large farms, for example, is due to farm 8 having a very large acreage of sugar beet. The relative accuracy of the two methods of estimation, apart from bias, is discussed in Example 7.17.

In the Survey of Fertilizer Practice the second method of estimation was used in investigation of secondary points, *e.g.* comparison of different types of farms. For the more important estimates, such as mean dressings per acre, a modification of the first method was used, the total acreages of sugar beet on the farms being taken as the raising factors for the second stage. This method of estimation is slightly more accurate than the first method given above (the actual gain in accuracy is discussed in Section 8.9), but will be biased if there is any tendency for farmers to apply heavier (or lighter) dressings to their large fields. There is no evidence that any appreciable bias does in fact arise from this cause, but even so it is perhaps doubtful whether there is much advantage in using this method of estimation rather than the unbiased method given above. The method would have been unbiased had selection of fields within a farm been made proportional to area, but this would have demanded somewhat more elaborate methods of selection in the field.

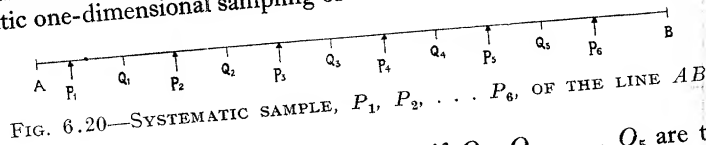
6.20 Systematic and balanced samples

The methods of estimation described in the preceding sections are also appropriate for systematic and balanced samples. Samples of these types without other restrictions, for instance, can be treated as if they were random samples for the estimation of the population values (but *not* for the sampling errors); if there is stratification the procedure for stratified random samples holds. An example of a systematic stratified sample with variable sampling fraction has already been given (Example 6.7).

Certain estimation processes are naturally inappropriate to systematic and balanced samples. If the process of selection in a systematic sample is such, for example, that stratification is automatically introduced, there will be no gain from stratification after selection. Equally in a balanced sample the variate for which balance has been effected will be of no further value as supplementary information—the balance ensures that the corrections based on regression, or ratio, will be zero, whatever the value of the regression coefficient. If each stratum is balanced separately, then the corrections for the different strata will all be zero, even if the regression coefficient or ratio varies from stratum to stratum.

In systematic samples of material which varies in a continuous manner, some gain in accuracy may result from the use of what are known as

end-corrections. These corrections are made by assigning to the boundary observations weights which depend on their distance from the boundary. In systematic one-dimensional sampling of a line AB (Fig. 6.20), for example, with



sampling points located at $P_1, P_2, \dots, P_6$, if $Q_1, Q_2, \dots, Q_5$ are the mid-points of P_1P_2 , etc., we may regard the observations at P_2, P_3, P_4, P_5 as estimates covering the lengths Q_1Q_2, Q_2Q_3 , etc. The observations at P_1 and P_6 can similarly be regarded as estimates covering the lengths AQ_1 and Q_5B . Consequently the weights assigned to P_1 and P_6 , relative to that assigned to P_2, P_3 , etc., will be AQ_1/Q_1Q_2 and Q_5B/Q_4Q_5 .

The same principle, or an adaptation of it, might be applied in the case of two-dimensional systematic sampling of areas. End-corrections, however, are not likely to be of much value in the type of material usually dealt with in census and survey work, and we shall not discuss them further here, beyond mentioning that if the regions near the boundary differ from the remainder of the area, the use of end-corrections, instead of separate treatment of the boundary regions in the manner outlined in Section 3.14, will lead to biased estimates.

6.21 Sampling on two successive occasions

The most straightforward procedure for estimating the values of the population mean on two successive occasions is to treat each occasion separately, following whatever method of estimation is appropriate to the sample obtained on that occasion, regardless of the values obtained on the other occasion. Such estimates may be termed *overall* estimates.

With independent samples on the two occasions, or with the same fixed sample on each occasion, the overall estimates will contain virtually all the available information, but where a sub-sample is taken on the second occasion, or there is partial replacement of the sample on the second occasion, the situation is more complicated.

If the sample on the second occasion is confined to a sub-sample of the original sample, change will be most simply estimated from the differences of the units included in the sub-sample only. An estimate of the population mean or total on the second occasion is similarly obtained by adding the estimated change to the overall estimate on the first occasion. The most accurate estimate of the population mean on the second occasion will, however, be obtained by calculating the regression of the sample values for the second occasion on the corresponding values for the first occasion, and using the sample values for the first occasion as supplementary information. The procedure is exactly the same as has been already outlined for use of

supplementary information by the regression method, the method appropriate to the type of sampling used being followed. The most accurate estimate of the change can then be obtained by taking the difference of this estimate of the mean from the overall estimate for the first occasion.

The formulæ for this procedure are as follows. Denote the sample values on the first occasion by x and those on the second occasion by y , the values belonging to units included in both samples by x' , y' , and those included on the first occasion only by x'' . If a fraction λ of all the units included on the first occasion are taken on the second occasion, and a fraction μ equal to $1 - \lambda$ are omitted, then, for a random sample,

$$\begin{aligned}\bar{x} &= \lambda \bar{x}' + \mu \bar{x}'' \\ \bar{y} &= \bar{y}' + b(\bar{x} - \bar{x}') \\ &= \bar{y}' + \mu b(\bar{x}'' - \bar{x}')\end{aligned}$$

where $\bar{x}$ is the overall estimate for the first occasion and $\bar{y}$ the adjusted estimate for the second occasion. The change is consequently estimated by

$$\bar{y} - \bar{x} = \bar{y}' - \bar{x}' - \mu(1 - b)(\bar{x}'' - \bar{x}')$$

The calculation of the regression coefficient is based on the values of the units which are included on both occasions.

If the changes of individual units are small compared with the differences between units, *i.e.* if the correlation between units on the two occasions is very close, as is likely to be the case when this type of sampling is adopted, b will be nearly equal to unity, and the estimate of change will differ little from that, $\bar{y}' - \bar{x}'$, derived from the units included on both occasions only. Equally the estimate of the population mean or total will differ little from that obtained by adding the estimate of change to the overall estimate on the first occasion.

When a sample of the same size is taken on each occasion with partial replacement, a fraction μ of the units being replaced and a fraction λ being retained, the sample units which are retained can be used to furnish an estimate $\bar{y}_1$ of the population mean on the second occasion by the regression method already given. In addition there will be a further independent estimate $\bar{y}_2$, equal to $\bar{y}'$, derivable from the sampling units which are included on the second occasion only. The most accurate estimate $\bar{y}_w$ will be provided by a weighted mean of these two estimates. The correct weights are $\lambda/(1 - \mu^2 r^2)$ and $\mu(1 - \mu r^2)/(1 - \mu^2 r^2)$, where r is the *correlation coefficient* between the unit values on the first and second occasions.

The correlation coefficient is calculated in the same manner as the regression coefficient b , using the values of the units common to both occasions, with the exception that instead of dividing by a quantity of the type $S(x - \bar{x})^2$ we divide by the corresponding quantity of the type $\sqrt{\{S(x - \bar{x})^2 S(y - \bar{y})^2\}}$. Thus, for a pair of random samples,

$$r = \frac{S(x' - \bar{x}')(y' - \bar{y}')}{\sqrt{\{S(x' - \bar{x}')^2 S(y' - \bar{y}')^2\}}}$$

In the more complicated types of sampling the sums of squares and products are modified in the same manner as in the calculation of b .

We thus have

$$\begin{aligned}\bar{y}_w &= \frac{\lambda}{1 - \mu^2 r^2} \bar{y}_1 + \frac{\mu(1 - \mu r^2)}{1 - \mu^2 r^2} \bar{y}_2 \\ &= \frac{\lambda}{1 - \mu^2 r^2} \{\bar{y}' + b(\bar{x} - \bar{x}')\} + \left(1 - \frac{\lambda}{1 - \mu^2 r^2}\right) \bar{y}''\end{aligned}$$

If the numbers in the sample on the two occasions are not the same the above formula takes the modified form

$$\bar{y}_w = \frac{n' \{\bar{y}' + b(\bar{x} - \bar{x}')\} + n''(1 - \mu r^2) \bar{y}''}{n' + n''(1 - \mu r^2)} \quad (6.21.a)$$

where n' is the number of units re-sampled on the second occasion, n'' is the number of new units, and μ is the proportion of units sampled on the first occasion which are not re-sampled on the second occasion.

An estimate of the change can similarly be obtained by taking the weighted mean of the two estimates, $\bar{y}' - \bar{x}'$ and $\bar{y}'' - \bar{x}''$. The weights to be assigned to these estimates are $\lambda/(1 - \mu r)$ and $\mu(1 - r)/(1 - \mu r)$, so that

$$\text{Change} = \frac{\lambda}{1 - \mu r} (\bar{y}' - \bar{x}') + \frac{\mu(1 - r)}{1 - \mu r} (\bar{y}'' - \bar{x}'') \quad (6.21.b)$$

This estimate of the change will differ from that given by the difference of $\bar{y}_w$ and the overall estimate on the first occasion. The reason for this is that once the sample for the second occasion has been taken, a more accurate estimate of the population mean on the first occasion is possible by using the information provided by the sample on the second occasion as supplementary information. If this revised estimate $\bar{x}_w$ is calculated, then the estimate of change given above will be very nearly equal to $\bar{y}_w - \bar{x}_w$. The slight discrepancy arises from the fact that unless the variances on the two occasions are equal the estimate of change given above is not quite the most accurate possible.

It will be noted that when r is equal to 1, the above estimate of change is equal to $\bar{y}' - \bar{x}'$, whereas if r equals 0 the estimate is equal to the difference of the overall estimates of the population means. Similarly, if r equals 0, $\bar{y}_w$ equals the overall mean of y , and if the values of each unit are the same on both occasions ($\bar{x}' = \bar{y}'$, $r = 1$, $b = 1$), $\bar{y}_w$ equals the mean of all the sample values included on each occasion, each value being included once only.

A further practical point arises in connection with the estimation of b . If the variability on the two occasions is the same, both the regression coefficients, y on x and x on y , will be equal to the correlation coefficient. In most material for which sampling on successive occasions of the type under consideration is likely to be used, the variability on the different occasions may be expected to be very similar, and for such material it is best to replace b by r , as the latter is less subject to errors of estimation.

Example 6.21

The percentage solids-not-fat in two successive months for all the 16 cows in a herd which were in their 2-6 months of lactation in one, at least, of these months were observed to be :

Cow	1	2	3	4	5	6	7	8
November	8.82	8.94	9.86	8.90	9.00	9.13	8.90	9.02
December	—	—	—	—	8.98	8.66	8.68	8.86
Cow	9	10	11	12	13	14	15	16
November	9.46	9.52	9.28	9.22	—	—	—	—
December	9.30	9.50	9.13	9.32	9.38	8.78	9.10	9.04

Estimate the change in percentage solids-not-fat between the two months, the mean percentage for December and the revised mean percentage for November.

The sampling is here due to natural causes, and the number of cows in their 2-6 months of lactation will therefore not remain completely constant. The above two months were selected from more extensive records.

We find :

$$\begin{aligned}
 S(x') &= 73.53 & S(y') &= 72.43 \\
 S(x'') &= 36.52 & S(y'') &= 36.30 \\
 \bar{x}' &= 9.1912 & \bar{y}' &= 9.0538 & \bar{y}' - \bar{x}' &= -0.1374 \\
 \bar{x}'' &= 9.13 & \bar{y}'' &= 9.075 & \bar{y}'' - \bar{x}'' &= -0.055 \\
 \bar{x} &= 9.1708 & \bar{y} &= 9.0608 & \bar{y} - \bar{x} &= -0.11 \\
 S(x' - \bar{x}')^2 &= 0.3435 & S(x' - \bar{x}')(y - \bar{y}') &= 0.4076 & S(y' - \bar{y}')^2 &= 0.6742
 \end{aligned}$$

$$r = 0.4076 / \sqrt{(0.3435 \times 0.6742)} = 0.847$$

It will be noted that the value of b is greater than unity, which illustrates the point made above that r provides a better estimate of the regression in material of this kind. The estimation of r should normally be based on more extensive data, though no very high accuracy is required—100 pairs of values will be fully adequate. In the present case the more extensive data confirm that the above value of r is about correct.

We also have $\lambda = \frac{2}{3}$, $\mu = \frac{1}{3}$, and thus

$$\begin{aligned}
 \frac{\lambda}{1 - \mu^2 r^2} &= 0.724 & \frac{\mu(1 - \mu r^2)}{1 - \mu^2 r^2} &= 0.276 \\
 \frac{\lambda}{1 - \mu r} &= 0.929 & \frac{\mu(1 - r)}{1 - \mu r} &= 0.071
 \end{aligned}$$

$$\begin{aligned}
 \bar{y}_w &= 0.724 \{9.0538 + 0.847(9.1708 - 9.1912)\} + 0.276 \times 9.075 = 9.0471 \\
 \text{Change} &= 0.929(-0.1374) + 0.071(-0.055) = -0.1315
 \end{aligned}$$

The estimate of the mean for November, revised on the basis of the December values, is

$\bar{x}_w = 0.724 \{9.1912 + 0.847 (9.0608 - 9.0538)\} + 0.276 \times 9.13 = 9.1786$
 giving the check, $9.0471 - 9.1786 = -0.1315$. Agreement is here exact, since the two regression coefficients, y on x and x on y , have both been taken equal to r .

6.22 Sampling on a number of successive occasions

The formulæ of estimation given in the last section cover all cases of sampling on two occasions only. When sampling is carried out with partial replacement on more than two occasions no such simple general solution is possible, but certain approximate solutions, which are very similar in form to those for sampling on two occasions, are likely to be sufficient for most practical purposes.

In a sampling scheme which is repeated at intervals it is generally desirable to provide as accurate an estimate as possible of the population mean on each occasion without any revision of the estimates for previous occasions. Suppose that $\bar{y}_h$ is the most accurate estimate which can be obtained for occasion h , taking into account the results of the sampling up to and including this occasion h , and that $\bar{y}_{h-1}$ is the similar estimate for occasion $h-1$, taking into account the results up to and including occasion $h-1$. Subject to certain limitations, $\bar{y}_h$ and $\bar{y}_{h-1}$ are related by a formula of the type

$$\bar{y}_h = (1 - \varphi) \{\bar{y}_h' + r(\bar{y}_{h-1} - [\bar{y}_{h-1}'])\} + \varphi \bar{y}_h'' \quad (6.22.a)$$

where suffices indicate the occasion, single dashes units common to occasions h and $h-1$, the mean on the earlier occasion being distinguished by square brackets, and double dashes units occurring on occasion h only.

The limitations are that a given fraction of the units is replaced on each occasion, that the variability on the different occasions and the correlation r between successive occasions are constant, and that the correlation between occasions two apart is r^2 , that between units three apart is r^3 , etc. This last condition is only necessary when units are included for more than two occasions, and no great loss of accuracy will occur under normal circumstances if it does not hold exactly.

The value of φ depends on the value of r , on the fraction μ replaced on each occasion, and on the number of occasions h on which samples have already been taken. With increasing h , φ rapidly tends to a limiting value, which depends only on r and μ . This limiting value is

$$\varphi = \frac{-(1 - r^2) + \sqrt{(1 - r^2)\{1 - r^2(1 - 4\lambda\mu)\}}}{2\lambda r^2}$$

The values for $h=2$ have been given in the previous section. For practical purposes the limiting value of φ may be used for all occasions after the second (the above formula for φ is due to Mr. H. D. Patterson).

When the value of r has been determined, the values of φ can be calculated and formula 6.22.a used.

For most practical purposes $\bar{y}_h - \bar{y}_{h-1}$ will provide an adequate estimate of the change between occasions $h-1$ and h . If change is of particular interest, however, formula 6.21.b may be used.* This latter estimate will of course not agree exactly with $\bar{y}_h - \bar{y}_{h-1}$ and will therefore lead to apparent inconsistencies in the summary of the results.

It sometimes happens that the sampling scheme, though broadly following a partial replacement procedure, gives rise to some inequality of numbers of units on the different occasions. This can be allowed for by substituting for φ the value φ' given by

$$\varphi' = \frac{n_h''}{\mu n_h} \varphi \quad (6.22.b)$$

where n_h is the number of units on occasion h , and n_h'' is the number of units not included on the previous occasion.

TABLE 6.22—PERCENTAGE SOLIDS-NOT-FAT: ADJUSTMENT OF SAMPLES TAKEN ON SUCCESSIVE OCCASIONS

	January	February	March	April	May	June
$\bar{y}_h$	5 9.400	11 9.090	9 9.111	10 9.059	8 9.211	11 9.345
$\bar{y}'_h$	—	4 9.288	8 9.122	8 9.086	3 9.060	7 9.326
$\bar{y}''_h$	—	7 8.977	1 9.020	2 8.950	5 9.302	4 9.380
$\{\bar{y}_h\}$	—	9.341	9.163	9.188	9.226	9.322
$\bar{y}_h$	9.400	9.122	9.151	9.152	9.262	9.338
$[\bar{y}'_h]$	4 9.335	8 9.072	8 9.025	3 8.947	7 9.267	—
$\bar{y}_h - [\bar{y}'_h]$	+ .065	+ .050	+ .126	+ .205	— .005	—
φ'	—	.603	.084	.151	.472	.275
From differences	9.400	9.353	9.403	9.464	9.577	9.636

* To obtain the most accurate possible estimate of change the information from occasions prior to $h-1$ would have to be taken into account. The procedure has been investigated by Mr. H. D. Patterson.

Example 6.22

Similar data on percentage solids-not-fat to those of Example 6.21 are given in abstract form in Table 6.22 for the months January-June. Only the 3-5 months of lactation are included. Obtain estimates of the mean percentage in successive months.

The table shows for each month the overall mean $\bar{y}_h$, the mean of cows occurring in the previous month $\bar{y}_h'$, and the mean of new cows $\bar{y}_h''$. The numbers of cows on which these means are based are also shown. The mean for the month $h-1$ of cows occurring in months h and $h-1$ is shown in the line $[\bar{y}_h']$ in the column for month $h-1$. Thus 9.335 and 9.288 are the means for January and February of the four cows occurring in both these months.

Summation of the sums of squares and products of deviations of pairs of entries for successive months from January to December gives an overall value for r of 0.811, so that r^2 is 0.657. The similar calculation of the correlation between months two apart gives r' equal to 0.746. The assumption that r' equals r^2 therefore somewhat underweights the information obtainable from occasions two apart.

The average value of μ over a long period will be $1/3$, but considerable fluctuations in numbers occur from month to month. For $\mu = 1/3$ the value of φ for occasions subsequent to the second is

$$\varphi = \frac{-(1 - 0.657) + \sqrt{(1 - 0.657)\{1 - 0.657(1 - 4.2/3.1/3)\}}}{2 \times 2/3 \times 0.657} = 0.252$$

Hence for March

$$\varphi' = \frac{1}{9/3} 0.252 = 0.084$$

etc. These values are shown in Table 6.22.

For the second occasion formula 6.21.a may be used. This is of the same form as formula 6.22.a, and gives

$$\varphi' = \frac{7(1 - 0.657/5)}{4 + 7(1 - 0.657/5)} = 0.603$$

Had the value for $\mu = 1/3$ been calculated, and corrected by means of formula 6.22.b, we should have obtained

$$\varphi = 1 - \frac{2/3}{1 - 0.657/9} = 0.281$$

$$\varphi' = \frac{7}{11/3} 0.281 = 0.536$$

which does not differ greatly from the correct value. Equally φ differs little from the value 0.252, obtained above for subsequent occasions.

The remainder of the calculations follow a standard pattern. The quantity $\{\bar{y}_h\}$, equal to $\bar{y}_h' + r(\bar{y}_{h-1} - [\bar{y}_h' - 1])$, is calculated, and the weighted mean of $\bar{y}_h''$ and $\{\bar{y}_h\}$ taken, with weights equal to ϕ' and $1 - \phi'$. Thus for February

$$\{\bar{y}_h\} = 9.288 + 0.811 \times (+0.065) = 9.341$$

$$\bar{y}_h = 0.603 \times 8.977 + 0.397 \times 9.341 = 9.122$$

The overall estimates $\bar{y}_h$ and the estimates from differences $\bar{y}_h' - [\bar{y}_h' - 1]$ are shown for comparison. The differences show a tendency to cumulative errors, which is to be expected even with close correlation.

It will be seen that once a value for r has been determined, and provision has been made to abstract the means $\bar{y}_h'$, $\bar{y}_h''$, and $[\bar{y}_h' - 1]$, the calculations are very simple, and can easily be undertaken for large-scale surveys, even when a number of different quantities require estimation.

ESTIMATION OF THE SAMPLING ERROR

7.1 Sampling errors of a random sample

The general principles involved in the estimation of sampling errors can best be made clear by considering the error of a random sample drawn from a large population.

Consider first a sample consisting of a single unit. Let the mean of the population be $\bar{y}$, and let the deviations of the individual values from this mean be $z_1, z_2, \dots$, so that $z_1 = y_1 - \bar{y}$, $z_2 = y_2 - \bar{y}$, $\dots$. Then the actual error in the estimate of the mean from a sample of one unit will be z_r , where r is the selected unit.

The mean of all the z 's is zero, and therefore the average of the errors of the estimates from a large number of samples of one unit (having regard to the signs of the errors) will approximate to zero. This is equivalent to saying there is no bias in the estimate.

In order to obtain a measure of the magnitude of the expected error we must therefore obtain some form of average of the z 's which does not take account of sign. One simple measure which might be taken is the average of all the z 's without regard to sign, but an alternative measure, which has a number of statistical advantages, is provided by the mean of the squares of all the z 's. This is termed the *mean square deviation* of y or the *variance* of y , and is denoted by $V(y)$ or σ^2 , and its estimate by $V(y)$ or s^2 . The square root of this variance is termed the *standard deviation* σ of a single unit.

In the same way, if a sample contains a number of units we may define the sampling variance of an unbiased estimate, say $\bar{y}$, derived from such a sample as the mean of the squares of the actual errors of a large number of samples of the same size. This variance will be denoted by $V(\bar{y})$, or, if estimated, by $V(\bar{y})$. The square root of this variance is generally termed the *standard error* of the estimate, and will be denoted by $S.E.(\bar{y})$, or, if estimated, by $S.E.(\bar{y})$. The term standard error is also sometimes applied to the standard deviation of a single unit, particularly when the deviations are in the nature of errors of observation.

The standard error of the estimate of the population mean derived from a sample of one unit is therefore equal to the standard deviation of a single unit. If a sample of two units r and s is taken, the actual error of the estimate of the population mean will be

$$\bar{y} - \bar{y} = \bar{y} - \bar{y} = \frac{1}{2}(z_r + z_s)$$

The standard error of the estimate will therefore be given by the square root of the average value of

$$\frac{1}{4}(z_r + z_s)^2 = \frac{1}{4}(z_r^2 + z_s^2 + 2z_r z_s)$$

If the population is large the average value of $z_r z_s$ is zero, as can be seen if

we consider a series of samples having the same first unit r and different second units s . The average values of z_r^2 and z_s^2 are both σ^2 . Hence the average value of the above expression is $\frac{1}{2}\sigma^2$. Consequently the standard error of the estimate is $\sigma/\sqrt{2}$.

It will be noted that the above argument does not depend on the form of the distribution of the z 's—there is no need, for example, for positive and negative deviations to be equally frequent. It does, however, require that each unit of the sample shall be randomly and independently selected. If, for instance, there were a tendency to select a second unit with a deviation similar to the first unit the average value of $z_r z_s$ would not be zero.

The argument can easily be extended to a sample of n units, for which the variance and standard error of the estimate will be found to be

$$V(\bar{y}) = \frac{1}{n} V(y), \quad \text{S.E.}(\bar{y}) = \frac{1}{\sqrt{n}} \sigma \quad (7.1.a)$$

We thus have the important general result that *the standard error of the estimate of the mean of a large population from a random sample is inversely proportional to the square root of the number of units in the sample.*

The standard error of the estimate of the total follows immediately from the rule that if l is any multiplier the standard error of $l\bar{y}$ is equal to l times the standard error of $\bar{y}$, provided l is not subject to sampling variation. Thus

$$\text{S.E.}(Y) = \text{S.E.}(gn\bar{y}) = gn \{\text{S.E.}(\bar{y})\} = g\sigma\sqrt{n} \quad (7.1.b)$$

Although $S(y)$ is not itself an estimate it is often convenient to consider its sampling variance or standard error. From the above rule

$$V\{S(y)\} = n V(y), \quad \text{S.E.}\{S(y)\} = \sigma\sqrt{n} \quad (7.1.c)$$

The standard error of an estimate can be expressed as a percentage of the population value of the estimated quantity. This form of expression is useful, as the percentage standard error is unaffected by the units in which the estimate is expressed, and the percentage standard error of the mean, of the total of the sample, and of the estimate of the population total are all equal. Similarly the standard deviation of a single unit can be expressed as a percentage of the mean value of a single unit. This is sometimes termed the *coefficient of variation*. Denoting it by $\sigma\%$, we have, in a large population,

$$\text{S.E.}\%(\bar{y}) = \text{S.E.}\% \{S(y)\} = \text{S.E.}\%(Y) = (\sigma\%)/\sqrt{n}$$

Thus in a population with a percentage standard deviation per unit of 20 per cent., the percentage standard error of the estimate of the population mean or total from a sample of 100 will be 2 per cent., that from a sample of 400 will be 1 per cent., etc.

In order to estimate the standard error of the mean or total in numerical terms an estimate of the value of σ will be required. This can be obtained from the deviations $y - \bar{y}$ of the numerical values of the selected units from their mean. $y - \bar{y}$ will be nearly, though not exactly, equal to z , and to a

first approximation an estimate of σ^2 will therefore be provided by the mean square deviation $S(y - \bar{y})^2/n$. Actually the sum of the squares of the deviations from the sample mean is always less than the sum of the squares of the deviations from the population mean, as can be seen from the identity

$$S(y - \bar{y})^2 = S(y - \bar{y})^2 - n(\bar{y} - \bar{y})^2$$

The average value of the first term on the right-hand side is $n\sigma^2$, and the average value of the second term is σ^2 , since $\bar{y} - \bar{y}$ is the error in the estimate of the mean. Thus $S(y - \bar{y})^2$ has an average value of $(n - 1)\sigma^2$, and consequently an estimate of σ^2 is given by

$$s^2 = \frac{1}{n - 1} S(y - \bar{y})^2 \quad (7.1.d)$$

The divisor $n - 1$ is technically known as the number of *degrees of freedom* associated with the estimate of error, and is equal to the number of independent comparisons that can be made between n values.

The calculation of the sum of the squares of the deviations $S(y - \bar{y})^2$ is best done from the sum of the squares of the values themselves. By this procedure the calculation and squaring of the individual deviations, which often involve fractional values, is avoided. One of the expressions

$$\begin{aligned} S(y - \bar{y})^2 &= S(y^2) - n\bar{y}^2 \\ &= S(y^2) - \bar{y} S(y) \\ &= S(y^2) - \frac{1}{n} \{S(y)\}^2 \end{aligned}$$

is used. The last term of each of the three expressions is usually termed "the correction for the mean." In calculating it from one of the first two expressions, $\bar{y}$ must be taken to at least as many significant figures as are required in the correction. For this reason the last expression is often the most convenient.

Sometimes it pays to use some convenient round number y_0 as a working mean, in which case we have

$$S(y - \bar{y})^2 = S(y - y_0)^2 - n(\bar{y} - y_0)^2$$

etc.

If a calculating machine is available the individual squares should not be written down—the sum of squares can be obtained directly by squaring the numbers successively without clearing the machine.

The calculation of the sum of squares from grouped data is illustrated in Examples 7.1.b and 7.2.b.

Example 7.1.a

Estimate the standard error of the estimate of the mean of the population (assumed large) of which the values of Table 6.4 are a sample.

The computations are as follows :

$$\begin{aligned}
 n &= 20 & S(y^2) &= 1959.12 \\
 S(y) &= 193.8 & \bar{y} S(y) &= 1877.92 \\
 \bar{y} &= 9.69000 & S(y - \bar{y})^2 &= 81.20 \\
 s^2 &= \frac{81.20}{19} = 4.274 = 2.07^2 \\
 \text{S.E.}(\bar{y}) &= \sqrt{\frac{4.274}{20}} = \pm 0.462
 \end{aligned}$$

Example 7.1.b

Table 7.1 gives the distribution of family income in a sample of 162 white families in Norfolk-Portsmouth, Virginia. Calculate the mean income of the sample and the sampling standard error of this mean.

TABLE 7.1—ANNUAL NET INCOME OF A SAMPLE OF 162 WHITE FAMILIES IN NORFOLK-PORTSMOUTH, VIRGINIA, 1934-6

Annual net income \$	No. of families	Working units	Calculation		Calculation by successive summation	
			Total (2) × (3)	Sum of squares (2) × (3) ² = (3) × (4)	Total	Sum of squares
(1)	(2)	(3)	(4)	(5)	(6)	(7)
600-	10	- 3	- 30	90	10	10
900-	23	- 2	- 46	92	33	43
1,200-	40	- 1	- 40	40	73	116
1,500-	32	0	0	0	116	
1,800-	28	+ 1	+ 28	28	57	105
2,100-	20	+ 2	+ 40	80	29	48
2,400-	4	+ 3	+ 12	36	9	19
2,700-	2	+ 4	+ 8	32	5	10
3,000-	1	+ 5	+ 5	25	3	5
3,300-	2	+ 6	+ 12	72	2	2
	162		- 11	495	105	358

With grouped data of this type it is best to use the group interval as the working unit and the central value of one of the central groups as the working mean. The group \$1500-1799 (central value, \$1649.5, since the data were rounded off to the nearest dollar before grouping) has been chosen. The calculation of the total and sum of squares in these units is shown in columns 4 and 5 of the table. The mean of the sample in working units is therefore $-11/162$, i.e. -0.06790 , and in the proper units is

$$1649.5 - 0.06790 \times 300 = 1629.1$$

The sum of squares of the deviations in the working units is

$$495 - 0.06790 \times 11 = 494.25$$

and in the proper units is therefore 494.25×300^2 , i.e. 44,482,000. Hence, dividing by 161, $s^2 = 276,290$, and the sampling standard error of the mean income is $\sqrt{(276,290/162)} = \pm 41.3$.

If no calculating machine, or only an adding machine, is available the alternative form of calculation shown in columns 6 and 7 may be preferred. Column 6 is formed from column 2 by successive summation from the ends. Column 7 is similarly formed from column 6. Note the check $73 + 57 + 32 = 162$ for column 6, and the checks of the final values for column 7 from the totals of column 6. The total in working units is then given by the difference of the totals of the two halves of column 6, i.e. by $105 - 116 = -11$. The sum of squares is obtained by doubling the total of column 7 and deducting the sum of the totals of the two halves of column 6, i.e. by $2 \times 358 - 105 - 116 = 495$.

7.2 Sampling from a finite population

The above theory requires modification in two respects if the population is not large. In the first place it is best to define σ^2 as

$$\sigma^2 = \frac{1}{N-1} S_p (y - \bar{y})^2$$

where S_p denotes summation over the whole population. This is equivalent to regarding the population as itself a random sample from an infinitely large population with variance σ^2 . With this definition of σ^2 formula 7.1.d for s^2 stands without modification. In certain textbooks and scientific papers the alternative definition with divisor N is adopted. This introduces the factor $(N-1)/N$ into the formula for s^2 , and leads to other complications in the discussion of the errors of sampling from finite populations which are avoided by the use of the first definition.

In the second place, the formula for the standard error of the estimate of the mean or other estimate requires modification by the introduction of the factor $\sqrt{1-f}$, or more strictly $\sqrt{1-f}$. Thus we have

$$\text{S.E.}(\bar{y}) = \sigma \sqrt{\frac{1-f}{n}} \quad (7.2)$$

That the introduction of some factor of this kind is necessary is obvious, since if the whole population is included in the sample ($f = 1$) the sampling error will be zero. The actual factor can be deduced by an extension of the algebraic analysis given above.

It should be noted that the factor $\sqrt{(1-f)}$ should not be introduced when testing the difference between the means of two sampled populations to see whether, for example, they are subject to different causal agencies. In this case we are concerned to determine whether there is a real and consistent difference running through all the units of the two populations: in other words, we wish to test whether the two samples can reasonably be regarded as random samples from a single infinitely large parent population, or whether they have to be regarded as samples of two different parent populations.

Example 7.2.a

Estimate the standard errors applicable to the estimates obtained in Example 6.4.b.

From Example 7.1.a, $s^2 = 4.274$, and consequently

$$\text{S.E.}(\bar{y}) = \sqrt{\left\{ \frac{4.274}{20} \left(1 - \frac{1}{25} \right) \right\}} = \pm 0.453$$

and since

$$\begin{aligned} Y &= 500 \bar{y} \\ \text{S.E.}(Y) &= 500 \times 0.453 = \pm 226 \\ \text{S.E.}(Y') &= 507 \times 0.453 = \pm 230 \end{aligned}$$

Example 7.2.b

Estimate the sampling error of the wheat acreage from the random sample of 125 farms of Table 6.6.a.

We find:

$$\begin{aligned} n &= 125 & S(y) &= 2301 & \bar{y} &= 18.4080 \\ S(y^2) &= 207,261 & S(y - \bar{y})^2 &= 164,904 & s^2 &= 164,904/124 = 1329.9 \\ \text{S.E.}(\bar{y}) &= \sqrt{\{1329.9(1 - \frac{1}{25})/125\}} = \pm 3.18 \\ \text{S.E.}(Y) &= 20\sqrt{\{125 \times 1329.9(1 - \frac{1}{25})\}} = 20 \times 125 \times 3.18 = \pm 7950 \end{aligned}$$

The calculation of the mean and of the sum of squares of deviations may alternatively be carried out by grouping the data. The groups should be so chosen that the distribution within any group containing a substantial number of values is reasonably even. A grouping interval of 10 acres is here convenient, but because of the large number of zeros these must be included in a separate group.

The grouped data and the calculation of the total and sum of squares are shown in Table 7.2. The calculations are carried out in terms of working

values in units of the grouping interval, and a working mean of 24.5 (the mean of group 4) is taken. The total is obtained from column 4 and the sum of squares from column 5. The mean in terms of the grouping interval is therefore

$$(-209.75 + 136)/125 = -73.75/125 = -0.590$$

and in terms of the proper units is

$$\bar{y} = 24.5 - 0.590 \times 10 = 18.60$$

Similarly

$$s^2 = (1719.2 - 73.75 \times 0.590) \times 10^2/124 = 167,570/124 = 1351.4$$

The rest of the computations proceed as before.

TABLE 7.2—CALCULATION OF THE MEAN AND VARIANCE FROM GROUPED DATA
(WHEAT ACREAGES OF TABLE 6.6.a)

Acres	Number	Working value	Total (2) × (3)	Squares (2) × (3) ²
0	80	- 2.45	- 196.00	480.2
1-	5	- 1.95	- 9.75	19.0
10-	4	- 1	- 4	4
20-	11	0	- 209.75	
30-	2	+ 1	+ 2	2
40-	2	+ 2	+ 4	8
50-	4	+ 3	+ 12	36
60-	4	+ 4	+ 16	64
70-	4	+ 5	+ 20	100
80-	3	+ 6	+ 18	108
90-	1	+ 7	+ 7	49
100-	3	+ 8	+ 24	192
110-	1	+ 9	+ 9	81
....				
260-	1	+ 24	+ 24	576
	125		+ 136	1719.2

If the data are fully tabulated, grouping is scarcely worth while for so small a body of data, even when a calculating machine is not available—especially when, as here, the existence of zeros complicates the grouping while simplifying the direct calculation of the sum of squares. With material on punched cards, however, the data can be most easily and compactly presented

in grouped form, and the advantages of this form of computation therefore become much greater, particularly when the number of values is large. Grouping also enables the form of the distribution to be much more easily comprehended. In the present data the relatively large number of values between 20 and 30, and the single very high value of 265, are immediately apparent.

7.3 The normal law of error

The above analysis shows that it is possible, from the numerical values of the selected sampling units, to estimate the standard error of the estimate of the mean of the population. This gives us a measure of the average error to be expected. The analysis has not, however, given us any indication of the frequency with which errors of different magnitudes may be expected to occur.

It is a matter of common observation that in most material which is subject to quantitative variation large deviations tend to occur less frequently than do small deviations. In much material, also, positive and negative deviations occur with about equal frequency. The exact distribution of the deviations of individual sampling units will, of course, vary considerably in different types of material, but it is a fortunate circumstance that, over a wide range of distributions of the parent material, the errors to which estimates such as the mean, total, etc., are subject are distributed approximately according to what is known as the *normal law of error*, i.e. in a *normal distribution*. Other things being equal, the larger the sample on which the estimate is based, the more closely is the law followed. If the deviations of the original material are normally distributed, the errors of the estimate of the mean, etc., will conform exactly to a normal distribution.

In a normal distribution the frequency with which deviations within the infinitesimal range z to $z + dz$ may be expected to occur is given by the expression :

$$f(z) dz = \frac{1}{\sigma\sqrt{2\pi}} e^{-z^2/2\sigma^2} dz$$

where σ is the *standard deviation*, and e is the base of Napierian logarithms, 2.71828 approximately.

Fig. 7.3 shows normal distributions with standard deviations $\sigma = 1$ and $\sigma = 2$. The vertical scale represents the frequency with which deviations within a range of 0.1 of z occur per 1000 values. Thus the ordinate at $z = 0$ for $\sigma = 1$ is 39.9, which indicates that on the average 39.9 values per 1000 may be expected to have deviates having values between -0.05 and $+0.05$. The value 39.9 can be derived from the above formula by putting $\sigma = 1$, $z = 0$, $dz = 0.1$ and multiplying by 1000.

From the figure it will be seen that positive and negative deviations of a given magnitude occur with equal frequency, and that large deviations are much less frequent than small ones. We are, however, in general not so much

concerned with the frequency of a deviation of any particular magnitude, as with the frequency with which deviations greater than a given magnitude may be expected to occur. These latter frequencies, which correspond to areas in Fig. 7.3, are shown in Table A.2 at the end of the book, for various values of z/σ . The area for $z/\sigma = 1$ is shaded for both curves.

From Table A.2 it will be seen that 61.7 per cent. of all values have a deviation or error (positive or negative) greater than one-half the standard deviation or standard error, 31.7 per cent. of all values have a deviation greater than the standard error, but only 4.6 per cent. have a deviation greater than twice the standard deviation, and only 0.27 per cent. will have a deviation greater than three times the standard deviation. Consequently, if we know

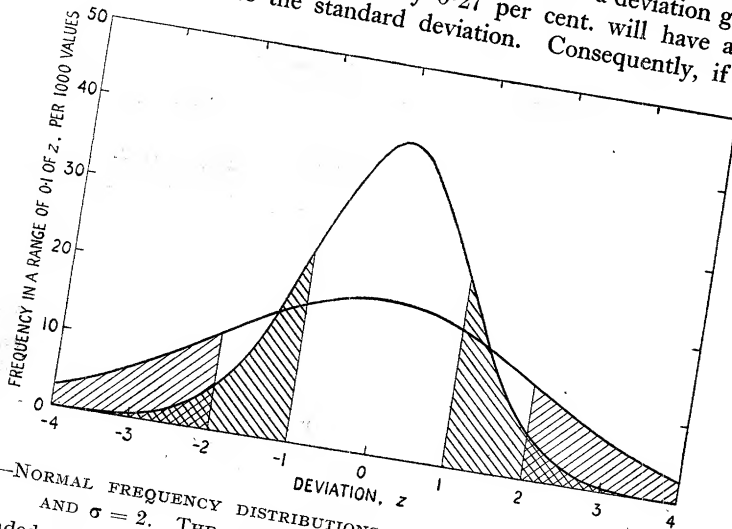


FIG. 7.3—NORMAL FREQUENCY DISTRIBUTIONS WITH STANDARD DEVIATIONS $\sigma = 1$ AND $\sigma = 2$. THE FLATTER CURVE IS THAT FOR $\sigma = 2$.

The shaded areas represent the total frequencies of the values for which the actual deviations are greater than the standard deviation.

that an estimate is subject to the normal law of error and has a given standard error, we can assign probable limits of error to this estimate. If limits of plus or minus twice the standard error are taken then in only 4.6 per cent. of the cases will the actual error lie outside these limits.

An alternative form of this statement, utilizing *fiducial probability*, which has certain logical advantages, is as follows.

If the true value of the mean is equal to the estimate minus twice the standard deviation (the lower limit of error) then a value of the estimate as high as, or higher than, that actually observed will be obtained in only 2.3 per cent. of all samples. (The values of Table A.2 are divided by 2 since deviations in only one direction are involved.) Similarly, if the true value of the mean is equal to the estimate plus twice the standard error (the upper limit of error), a value of the estimate as low as, or lower than, that actually observed will be

obtained in only 2.3 per cent. of all samples. For limits of plus or minus once the standard error the corresponding percentages are 16 per cent. These closer limits are useful as indicating the region within which, or in the fairly close neighbourhood of which, the mean is likely to lie.

As pointed out above, we usually only have an estimate of the standard error, which is itself subject to error, the accuracy being dependent on the number of degrees of freedom, and statements in the above form are therefore not exact.

The effect of inaccuracy in the estimate of error on fiducial statements can be allowed for by the use of what is known as the *t distribution*, instead of the normal distribution. In general, however, inaccuracies due to paucity of data are not sufficiently great for this to be necessary in census work. More important is the fact that with a number of types of sample frequently employed, *e.g.* systematic samples and stratified samples with one unit from each stratum, fully valid estimates of error are not available. The estimates of error actually obtained in such cases are usually overestimates of the sampling standard errors, and any exact fiducial statement is therefore impossible.

In a random sample of n ($n - 1$ degrees of freedom) from a normal distribution with standard deviation σ the standard error of s is given by

$$\text{S.E.}(s) = \frac{\sigma}{\sqrt{2(n-1)}} \quad (7.3)$$

If the material conforms approximately to the normal law of variation, an estimate based on 50 degrees of freedom will therefore determine the sampling error with a standard error of 10 per cent. and an estimate based on 200 degrees of freedom with a standard error of 5 per cent. If the material does not conform to the normal law the accuracy may be substantially less.

Example 7.3

Assign limits of error to the estimate of the mean of the population (assumed large) of the values of Table 6.4. (The values are actually a random sample from a normal distribution with mean 10 and standard deviation 2.) Show that they are distributed in the expected manner. Calculate also the standard error of the estimate of the standard deviation.

The results obtained in Example 7.1.a indicate that the true value of the mean is not likely to lie far outside the range 9.69 ± 0.46 , *i.e.* 9.23–10.15, and is fairly certain to lie within the range $9.69 \pm 2 \times 0.46$, *i.e.* 8.77–10.61.

The distribution of the 20 values is given in Table 7.3. Integral values have been allotted $\frac{1}{2}$ and $\frac{1}{2}$ to the two appropriate classes. Each interval in this grouping is equal to 0.5σ . From Table A.2 the proportionate frequency of observations with deviations greater than 0.5σ (positive or negative) is 0.6171, and consequently the expected frequencies of observations between 9 and 10 and between 10 and 11 are each $20 \times \frac{1}{2}(1 - 0.6171) = 3.83$.

TABLE 7.3—OBSERVED AND EXPECTED FREQUENCIES IN THE SAMPLE
OF TABLE 6.4 FROM A NORMAL DISTRIBUTION

Range	< 6	6-7	7-8	8-9	9-10	10-11	11-12	12-13	13-14	14-15	> 15	Total
Observed	0	1	2.5	5	2.5	5.5	1	0.5	1	1	0	20
Expected	0.45	0.88	1.84	3.00	3.83	3.83	3.00	1.84	0.88	0.33	0.12	20.00

Similarly the proportionate frequency of observations greater than $\pm 1.0\sigma$ is 0.3173, and consequently the expected frequencies between 8 and 9 and between 11 and 12 are each $20 \times \frac{1}{2} (0.6171 - 0.3173) = 3.00$. In this manner all the expected frequencies shown in Table 7.3 can be calculated. The observed frequencies conform satisfactorily to the expected frequencies.

From formula 7.3 the standard error of the estimate s is

$$\text{S.E.}(s) = \frac{2}{\sqrt{(2 \times 19)}} = \pm 0.324$$

The actual value of s , 2.07, is therefore closer to the true value than will occur on the average in samples of 20.

7.4 Qualitative variates

From the procedure developed for random samples it will be seen that the estimation of the sampling errors of estimates derived from a quantitative variate can be divided into two distinct stages, the first being the estimation of the variability of the individual sampling units (or more strictly the part of the variability which contributes to sampling error), and the second the derivation of the standard errors of the estimates in terms of the variability of the individual sampling units.

The same principles hold when the variate under consideration is qualitative. In the case of a random sample, however, the variability of an attribute of the sampling units depends only on the proportion of units possessing the attribute in the population. Hence in random samples no estimate of the variability of the individual sampling units is required.

For a random sample from a large population, if $q = 1 - p$, we have

$$V(p) = \frac{pq}{n}, \quad \text{S.E.}(p) = \sqrt{\frac{pq}{n}}$$

If the population is finite and the sampling fraction f is appreciable the formula becomes, to all necessary accuracy,

$$\text{S.E.}(p) = \sqrt{\{pq(1-f)/n\}}$$

The exact expression is obtained by replacing $(1-f)$ by $(N-n)/(N-1)$.

Similarly

$$V(u) = npq(1-f)$$

and hence

$$\text{S.E.}(U) = g\sqrt{\{npq(1-f)\}} = N\{\text{S.E.}(p)\}$$

Fig. 7.4 shows the way in which the standard error of the estimated percentage $S.E. (100 p)$ varies with the percentage $100 p$ in samples of 100 and 1000 from a large population. The actual values of $S.E. (100 p)$ are shown by the full line, while the dotted line gives the percentage standard error $100 S.E. (100 p)/100 p$.

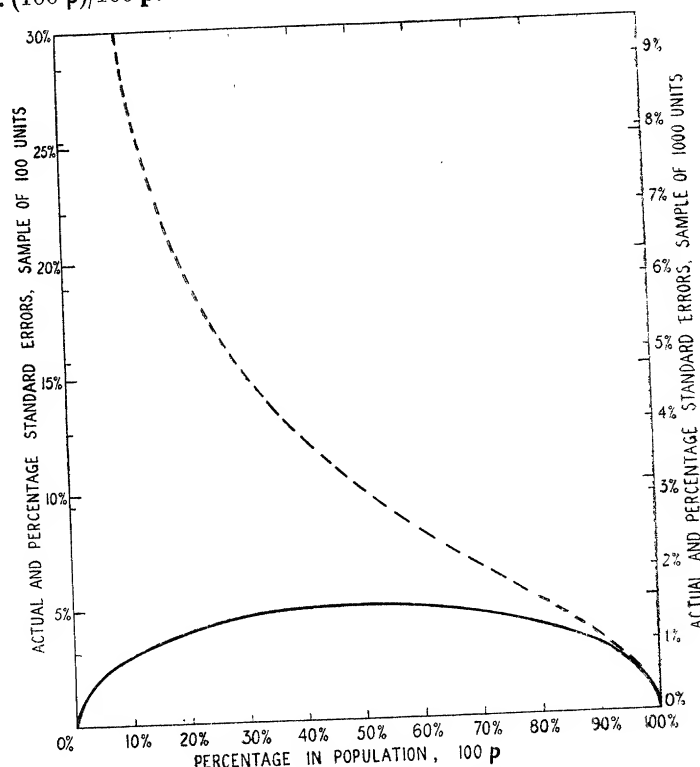


FIG. 7.4—STANDARD ERRORS OF THE ESTIMATED PERCENTAGE OF UNITS HAVING A GIVEN ATTRIBUTE, AND OF THE ESTIMATED NUMBER HAVING THE ATTRIBUTE, FOR DIFFERENT PERCENTAGES OF UNITS HAVING THE GIVEN ATTRIBUTE IN THE POPULATION

The full line shows the actual standard error of the estimated percentage, and the broken line the *percentage* standard error of the estimated number. This is equal to the percentage standard error of the estimated percentage. The scales shown are for samples of 100 and 1000. For a sample of 10,000 divide the values of the left-hand scale by 10, etc.

The standard errors obtained with larger samples for which the sample number is a power of 10 can also be read from the figure by dividing one of the scales by the appropriate power of 10. Thus for a sample of 10,000 the scale for the sample of 100 is divided by 10, since $\sqrt{(10,000/100)} = 10$.

The actual standard error has its maximum value at $p = 0.5$. At this point the standard error of the estimated percentage with a sample of 100 is 5.0, and with a sample of 1000 is 1.58, *i.e.* if the true percentage is 50 per

cent. the value of the estimated percentage will usually lie between 40 per cent. and 60 per cent. with a sample of 100, and between 47 per cent. and 53 per cent. with a sample of 1000. Expressed in percentage terms the standard errors and limits at this point are double the above values. As the percentage in the population decreases from 50 per cent. the actual standard error of the estimated percentage also decreases, but the percentage standard error continues to increase. With $100p = 20$ per cent. the actual standard error with a sample of 100 is 4.0 and the percentage standard error 20 per cent. ; with $100p = 5$ per cent. they are 2.2 and 44 per cent. respectively. Thus, while quite a small sample serves to verify that the proportion in a population possessing a given attribute is small, the determination with any accuracy of the actual number possessing the attribute requires a relatively large sample when the proportion is small.

In estimating the sampling error the proportion in the population p can be replaced by its estimate $\hat{p}$ from the sample, *i.e.* by the proportion in the sample. This results in a certain amount of error in the estimate of variability, since the proportion in the sample will not in general be exactly equal to that in the population, but in large samples, such as are commonly met with in census work, this is not likely to be of much importance. Exact treatment of the problem is possible by use of Table VIII.1 of *Statistical Tables for Biological, Agricultural and Medical Research*.

It must be clearly recognized that the above formulæ hold only when the units of which the proportion possessing a given attribute is being assessed are themselves the sampling units, and the sample is a random one from the whole population. In a stratified random sample the formulæ apply to each stratum taken separately. In other cases, *e.g.* multi-stage sampling, and all types of sampling with supplementary information, the variability no longer depends only on the proportions in the population or strata. Thus, for example, the formulæ are not applicable to the proportion of farms growing a given crop when two-stage sampling by administrative districts, and by farms within selected districts, has been carried out. Equally they are not applicable to the proportion of individuals of a given race in a human population, when the sampling has been by households ; since the whole of a household is usually of the same race, the variability will clearly be greater than if a random sample of individuals had been taken.

When the above formulæ do not hold, the variability of the individual sampling units must be assessed in the same manner as with a quantitative variate, scoring the qualitative variate 1 or 0. (See Example 7.8.b.)

Example 7.4

Estimate the sampling errors of the estimates of Example 6.4.a.

We have $p = 0.0738$, $q = 1 - 0.0738 = 0.9262$. Hence

$$\text{S.E.}(p) = \sqrt{\left[\frac{0.0738 \times 0.9262}{8491} \left(1 - \frac{1}{50} \right) \right]} = \pm 0.00281$$

$$\text{S.E. (percentage defective)} = 100 \times 0.00281 = \pm 0.281$$

$$\text{S.E. (total number defective)} = \text{S.E.}(U) = 50 \times 8491 \times 0.00281 = \pm 1190$$

Thus the percentage defective, 7.38 per cent., has a standard error of ± 0.28 , which implies, taking limits of plus or minus twice the standard error, that the true percentage defective probably lies between 6.8 per cent. and 7.9 per cent. Similarly the number defective probably lies between 29,000 and 33,700. Note that the standard error expressed as a percentage of the percentage defective or number defective, *i.e.* what is ordinarily called the percentage standard error, is

$$\text{S.E. \% (p)} = \text{S.E. \% (U)} = \frac{0.281}{7.38} \times 100 = \frac{1190}{31,350} \times 100 = \pm 3.8 \text{ per cent.}$$

These standard errors are likely to be slight overestimates, since the sample was in fact systematic.

7.5 Standard errors of functions of estimates

If we have a number of estimates $y_1, y_2, y_3, \dots$ with sampling errors which are independent, the sampling variances being $V(y_1), V(y_2), V(y_3), \dots$ and we form a linear function of the y 's:

$$L = l_1 y_1 + l_2 y_2 + l_3 y_3 + \dots$$

where the l 's are any multipliers whose values are not influenced by the sampling, the sampling variance of L is given by

$$V(L) = l_1^2 V(y_1) + l_2^2 V(y_2) + l_3^2 V(y_3) + \dots \quad (7.5.a)$$

The condition of independence is important. The sampling errors of two estimates will be independent if the estimates are derived from sets of values which are themselves independent. Estimates derived from samples of different populations, or from different strata of the same population, are consequently independent, as are estimates derived from different samples of a large population. Estimates derived from two different variates belonging to the same sampling units are not in general independent, since such variates are likely to be correlated, high values of the one being associated with high (or low) values of the other in the same sampling units.

A number of important simple formulæ are derivable from the above general formula.

The standard error of a multiple of an estimate is the same multiple of the standard error of the estimate:

$$\begin{aligned} V(l y_1) &= l^2 V(y_1) \\ \text{S.E.}(l y_1) &= l \text{S.E.}(y_1) \end{aligned} \quad (7.5.b)$$

This formula has already been used in Section 7.1.

The standard error of the difference of two independent estimates is the square root of the sum of the squares of the standard errors of the estimates:

$$\begin{aligned} V(y_1 - y_2) &= V(y_1) + V(y_2) \\ \text{S.E.}(y_1 - y_2) &= \sqrt{[\text{S.E.}(y_1)]^2 + [\text{S.E.}(y_2)]^2} \end{aligned} \quad (7.5.c)$$

The standard error of the sum of a number of independent estimates is the square root of the sum of the squares of the standard errors of the estimates:

$$V(y_1 + y_2 + y_3 + \dots) = V(y_1) + V(y_2) + V(y_3) + \dots \quad (7.5.d)$$

$S.E.(y_1 + y_2 + y_3 + \dots) = \sqrt{[S.E.(y_1)]^2 + [S.E.(y_2)]^2 + [S.E.(y_3)]^2 + \dots}$ which may be expressed by the rule that "variances are additive."

The standard error of the estimate of the mean of a large population can also be derived from the formula.

Weighted means are a type of linear function which occurs frequently in statistics. The general form of a weighted mean is

$$\bar{y}_w = \frac{w_1 y_1 + w_2 y_2 + \dots}{w_1 + w_2 + \dots}$$

where the w 's are the weights. Knowing the variances of the y 's, the variance of $\bar{y}_w$ can be calculated, provided the y 's are independent. Two cases are of frequent occurrence.

$$(1) V(y_1) = V(y_2) = \dots = V(y)$$

We then have

$$V(\bar{y}_w) = \frac{S(w^2)}{\{S(w)\}^2} V(y) \quad (7.5.e)$$

$$(2) V(y_1) = \lambda/w_1, \text{ etc., where } \lambda \text{ is a constant.}$$

We then have

$$V(\bar{y}_w) = \frac{S(w)}{\{S(w)\}^2} \lambda = \frac{\lambda}{S(w)} \quad (7.5.f)$$

This is the form of weighted mean which is used when we wish to obtain the most accurate combined estimate from a number of independent estimates of the same quantity whose relative variances are known. The weights are taken equal (or proportional) to the reciprocals of the variances, and the variance of the weighted mean is given by the reciprocal of the sum of the weights (or a multiple of this reciprocal).

A further type of weighted mean is that in which the weights w are in the nature of supplementary information, the quantities y and w both being determined from the individual sampling units, with the variances of the y 's related in some unknown manner to the w 's, and $\bar{y}_w = S(wy)/S(w)$.

In order to obtain an unbiased estimate of $V(\bar{y}_w)$, whatever the variance law, the squares of the deviations of the y from $\bar{y}_w$ must be weighted in proportion to w^2 before summation. For a random sample, if

$$Q = Sw^2(y - \bar{y}_w)^2$$

and

$$s_q^2 = Q/(n-1)$$

we have

$$V(\bar{y}_w) = \frac{(1-f)ns_q^2}{\{S(w)\}^2} \quad (7.5.g)$$

It may also be noted that if the variance of y for given w can be regarded as constant over the range of w , and there is also no variation in the mean value of y for given w over the range other than that ascribable to random variation in y , the *efficient* estimate of the variance of y is given by the ordinary formula

$$V(y) = S(y - \bar{y})^2 / (n - 1) \quad (7.5.h)$$

and formula 7.5.e may be used to estimate $V(\bar{y}_w)$, with the introduction of the factor $(1 - f)$. If the variance of y for given w is inversely proportional to w then the efficient estimate of the variance of y is given by

$$\left. \begin{aligned} Q' &= Sw(y - \bar{y}_w)^2 = Swy^2 - \bar{y}_w Swy \\ s_q'^2 &= Q' / (n - 1) \end{aligned} \right\} \quad (7.5.i)$$

$s_q'^2$ is an estimate of λ , and formula 7.5.f can be used for estimating $V(\bar{y}_w)$, with the introduction of the factor $(1 - f)$. Either of these estimates will be biased if the true variance law is different from that assumed or the other condition does not hold. They should therefore not be used without careful consideration.

The mean ratio $\bar{r}$ used in the ratio method of estimation is an example of a weighted mean of the above type, since $\bar{r} = S(y)/S(x) = S(xr)/S(x)$, and we therefore substitute r for y and x for w . This case is discussed in more detail in Sections 7.8–7.11, which deal with the estimation of errors in the ratio method in both random and stratified samples. Normally formula 7.5.g will be used to estimate $V(\bar{r})$, but under certain circumstances formulæ 7.5.h and 7.5.e may be employed.

The approximate formulæ for the standard errors of the product and the ratio of two estimates whose sampling errors are independent may also be noted. These are given by

$$V(y_1 y_2) = y_2^2 V(y_1) + y_1^2 V(y_2) \quad (7.5.j)$$

$$V\left(\frac{y_1}{y_2}\right) = \left(\frac{y_1}{y_2}\right)^2 \left(\frac{V(y_1)}{y_1^2} + \frac{V(y_2)}{y_2^2}\right) \quad (7.5.k)$$

These formulæ are only satisfactory if $V(y_1)$ and $V(y_2)$ are small relative to y_1^2 and y_2^2 respectively.

If the estimates $y_1, y_2, y_3, \dots$ are not independent the concept of covariance must be introduced. The covariance between two estimates is the mean product deviation, and is estimated in exactly the same manner as is the variance of each of the estimates, with the exception that the sum of squares of the deviations of a single variate is replaced by the sum of products of the deviations of the two variates. If the covariance between y_1 and y_2 is denoted by $\text{cov}(y_1 y_2)$ the additional terms

$$+ 2l_1 l_2 \text{cov}(y_1 y_2) + 2l_1 l_3 \text{cov}(y_1 y_3) + 2l_2 l_3 \text{cov}(y_2 y_3) + \dots$$

must be introduced into the formula for $V(L)$. This gives the additional term $-2 \text{cov}(y_1 y_2)$ in the formula for $V(y_1 - y_2)$. The corresponding

additional term in $V(y_1y_2)$ is $+2y_1y_2 \text{ cov}(y_1y_2)$, and that in $V(y_1/y_2)$ is $-2 \text{ cov}(y_1y_2)/y_1y_2$ within the bracket.

If $y_1, y_2, y_3, \dots$ are derived from different variates belonging to the same sampling units, *e.g.* measurements of different characters, the variance of any linear function L can, if desired, be estimated directly by calculating a value L for each sampling unit separately and estimating $V(L)$ from these values in the manner appropriate to a single variate. This obviates the calculation of the variances and covariances of the individual variates. The same method can be followed with products and ratios, subject to the same limitations as those given above for formulæ 7.5.e and 7.5.f. If the errors of a number of functions are required, however, it is best to calculate the variances and covariances (Example 7.8.b).

The regression and correlation coefficients can be expressed in terms of the variances and covariance. We have the relations $b = \text{cov}(xy)/V(x)$, and $r = \text{cov}(xy)/\sqrt{\{V(x) \cdot V(y)\}}$.

In the more complicated types of sampling, discussed later, the estimation of covariance is again exactly parallel to the estimation of the corresponding variances, the squares being replaced by products wherever they occur.

Example 7.5

Calculate standard errors for the various estimates of the regional and varietal differences between the yields of potatoes given in Tables 5.23.c, 5.23.d, 5.23.e and 5.24, given that the variance of the yield per acre of any one variety in any one region is 4.22, and that the standard deviation is therefore ± 2.05 .

The standard errors of the regional-varietal means of Table 5.23.b are obtained by dividing the above standard deviation by the square roots of the numbers of fields. Thus for Majestic in Scotland the standard error is $2.05/\sqrt{37} = \pm 0.34$. The standard errors are shown in Table 7.5.

TABLE 7.5—POTATO SURVEY: STANDARD ERRORS OF REGIONAL-VARIETAL MEANS

	Scotland	North	E. Midlands	South	West
Majestic . . .	± 0.34	± 0.24	± 0.20	± 0.20	± 0.24
King Edward . .	± 0.32	± 0.55	± 0.22	± 0.25	± 0.31
Great Scot . . .	± 0.48	± 0.55	—	± 0.84	± 0.48
Arran Banner . .	± 0.72	± 0.33	—	± 0.68	± 0.38
Kerr's Pink . . .	± 0.25	± 0.34	—	—	± 0.57

These standard errors enable the differences between the individual means to be examined more critically. The difference between Scotland and the Northern region for Arran Banner, for example, is at first sight anomalous, being -0.12 . The standard error of this difference is $\sqrt{(0.72^2 + 0.33^2)} = \pm 0.79$. This difference, therefore, does not conflict very seriously with the other differences.

On the other hand the difference between this difference and the largest positive difference, that for King Edward, is $+1.94 - (-0.12) = +2.06$. This quantity has a standard error of $\sqrt{(0.32^2 + 0.55^2 + 0.72^2 + 0.33^2)} = \pm 1.02$. It might therefore be judged unlikely, on this evidence alone, that the difference has arisen by chance, since Table A.2 shows that a difference of 2.0 times its standard error would arise by chance in less than 1 in 20 times. In statistical terminology the difference is *significant* at the 1 in 20 level of significance. This conclusion, however, is subject to the qualification that we have here picked the two extreme differences out of 10 possible pairs. A combined test* of all 5 differences shows that they are not exceptionally variable. A more comprehensive test of the differences of the whole table confirms this verdict. It may be noted, however, that Arran Banner is only $\frac{1}{3}$ as common in Scotland as in the Northern region, whereas King Edward is 3 times as common in Scotland as in the Northern region. The observed differences are therefore in the direction that would be expected if the varieties were grown in the regions to which they were most suited.

The standard errors of the means of Table 5.23.c may be calculated in a similar manner, that for mean (a) of Scotland for example being $\frac{1}{5}\sqrt{(0.34^2 + 0.32^2 + 0.48^2 + 0.72^2 + 0.25^2)} = \pm 0.20$. The corresponding value for the Northern region is ± 0.19 . The standard error of the estimated difference $8.55 - 7.46 = +1.09$ is therefore $\sqrt{(0.20^2 + 0.19^2)} = \pm 0.28$. Similarly the standard error for the corresponding difference of the means (b), $8.98 - 7.37 = +1.61$, is ± 0.38 .

Table 5.23.d provides an example of a weighted mean with the weights so chosen that the most accurate combined estimate is obtained. Formula 7.5.f is therefore appropriate, and λ represents the variance of a single field, *i.e.* $\lambda = 4.22$. Hence the variance of the weighted mean difference $= 4.22/74 = 0.0570$, and the standard error is therefore ± 0.24 .

The *relative efficiency* of the above estimates of the differences may be assessed from the ratio of the reciprocals of the variances (Section 8.1). Assigning a value of 100 to the weighted mean, the relative efficiencies of means (a) and (b) are 73.5 and 39.9.†

Finally we may evaluate the standard errors of Table 5.23.e. These cannot be evaluated exactly, as the pooling of regions is based on the assumption that there are no differences of any importance between these regions. In so

* The weighted sum of squares of deviations gives $\chi^2 = 5.74$ with 4 degrees of freedom.

† These values do not represent the true efficiencies, which are obtained by assigning the value of 100 to the most efficient possible method, here that of Table 5.24.

far as this is not the case additional errors will be introduced, and the standard errors calculated on the assumption that there are no differences will therefore be underestimates of the true errors.

The standard errors of the means of the pooled regions can be calculated from the numbers of fields on which each mean is based. Thus that for Majestic is $\sqrt{(4 \cdot 22/356)} = \pm 0.11$. The standard errors of the Scottish means have already been given in Table 7.5. The standard error of the weighted mean for Majestic is therefore given by

$$\sqrt{\left(\frac{4^2 \times 0.11^2 + 1^2 \times 0.34^2}{5^2} \right)} = \pm 0.11$$

Similarly the standard errors for the other four varieties are found to be ± 0.13 , ± 0.29 , ± 0.24 and ± 0.24 .

The standard errors of the estimates obtained in Table 5.24 can only be calculated exactly by inversion of the matrix of the simultaneous linear equations giving the least-squares solution. This requires a good deal of arithmetical work. The method is explained, for example, in *Statistical Methods for Research Workers*, Section 29.

In material of this type, however, there will rarely be any need to determine the standard errors exactly. An upper limit to the standard error of any particular difference can be obtained by calculating the standard error of the estimate given by Method (3) of Section 5.23. A lower limit can be obtained by calculating what the standard error would be if there were no cross classification, and if the relevant variance per unit were that within sub-classes. The value of this latter standard error for the difference between Scotland and the Northern region, for example, is

$$2.05 \sqrt{\left(\frac{1}{174} + \frac{1}{177} \right)} = \pm 0.22$$

The value of the standard error for Method (3) has already been found to be ± 0.24 . In this case close limits are set to the true standard error.

7.6 Stratified random sample with possibly unequal variances within strata

Since in a stratified sample differences between the sampling units in the different strata are eliminated from the sampling error, in estimating this error we require not the total variance of the sampling units over the whole population, but the variances of the sampling units within the different strata.

A simple example will illustrate the difference. Suppose we have a large population of which 25 per cent. of the units have the value 8, 50 per cent. have the value 10, and 25 per cent. have the value 12. The mean of the population is 10, and 50 per cent. of the values have a deviation of 0 from the mean, the remaining 50 per cent. having a deviation of ± 2 . The mean square deviation or total variance is therefore $0.5 \times 0 + 0.5 \times 2^2 = 2$.

Suppose now the population is divided into two strata, the first containing all the units with value 8 and one-half the units with value 10, the second one-half the units with value 10 and all the units with value 12. The mean of the first stratum is 9, and all units in it have a deviation of ± 1 . The mean square deviation or variance within the first stratum is therefore 1. The same holds for the second stratum.

In this example the strata are of the same size and the within-strata variances are equal. Examples can easily be constructed in which this is not the case, but even if the variances are unequal an average within-strata variance can be calculated, and in a large population this will always be less than the total variance if there are differences between the strata means.

If the sample numbers in all strata are sufficiently large for the within-strata variances per sampling unit to be separately estimated, the sampling variances of the means or totals of the individual strata can be estimated separately, and the sampling variance of the population mean or total, which is a linear function of these means or totals, can be obtained by the use of the formulæ of Section 7.5.

This method is valid even if there is inequality in the within-stratum variance per sampling unit from stratum to stratum, and is applicable to all types of stratification, including stratification with a variable sampling fraction and stratification after selection.

In general it is best to build up the variance of the population estimate under consideration by calculating the variances of the component parts, and adding these variances, or the correct multiples of them, the same steps being followed as in the calculation of the estimate itself. We will therefore not give formulæ for the variances of all the different estimates set out in Sections 6.4 and 6.5, but will illustrate the derivation of such formulæ by obtaining that for $V(\bar{y})$ in the case of a variable sampling fraction.

We have $\bar{y} = \Sigma\{g_i S_i(y)\}/N$. If σ_i^2 is the variance within the i th stratum, $V\{S_i(y)\} = n_i \sigma_i^2 (1 - f_i)$, and hence by formula 7.5.a

$$V(\bar{y}) = \Sigma\{g_i^2 n_i \sigma_i^2 (1 - f_i)\}/N^2 \quad (7.6.a)$$

For $V(\bar{y})$ the σ_i^2 will be replaced by their estimates s_i^2 .

If all the sampling fractions are equal we have, since $N = gn$,

$$V(\bar{y}) = (1 - f) \Sigma(n_i s_i^2)/n^2 \quad (7.6.b)$$

If we require the sampling errors of estimates applicable to domains of study which cut across the strata the situation is more complicated. If a dash is used to denote the domain in question, and if $s_i'^2$ is the estimated variance per unit between the units of this domain in stratum i , and the proportion of the selected units of stratum i not in the domain is q_i' , so that $q_i' = (n_i - n_i')/n_i$, estimates of the variances of the total and mean of the domain are given by

$$\begin{aligned} V(Y') &= \Sigma g_i'^2 n_i' (1 - f_i) \{q_i' \bar{y}_i'^2 + s_i'^2\} \\ N'^2 V(\bar{y}') &= \Sigma g_i'^2 n_i' (1 - f_i) \{q_i' (\bar{y}_i' - \bar{y}')^2 + s_i'^2\} \end{aligned}$$

Consequently, in the case of a stratified sample with uniform sampling fraction, an approximate estimate of the sampling error of a domain mean

will be obtained by treating the sample as if it were stratified for the domain in question, and not stratified for the strata which cut across the domain. (See Example 7.7.b.)

In the case of a variable sampling fraction the above formulæ must be used. In addition the variance of the difference of the means of two domains will be somewhat increased by covariance. It may be noted that, although the variances of estimates for such domains may be considerably increased by lack of control by stratification, a variable sampling fraction will still be advantageous in the appropriate circumstances. The optimal sampling fractions will be dependent on the estimates required, being approximately proportional to the square roots of $(1 - q_i')$ times the quantities in the curly brackets. The examples which follow will illustrate the details of the methods to be followed in the cases that are ordinarily met with in practice.

Example 7.6.a

Estimate the sampling errors of the estimates of Example 6.5.

The computations for acreages are set out in Table 7.6.a. The various steps are as follows.

The sums of squares of deviations from each stratum mean, $S_i(y - \bar{y}_i)^2$, are first calculated from the sample values given in Table 6.5.a, using the method of Example 7.1.a. The estimated within-strata variances s_i^2 are then calculated by dividing by $n_i - 1$. Multiplying these by $(1 - f)/n_i$ gives

TABLE 7.6.a—ESTIMATION OF SAMPLING ERRORS OF THE WHEAT ACREAGES OF EXAMPLE 6.5

Size-group	n_i	$n_i - 1$	$S_i(y - \bar{y}_i)^2$	s_i^2	$V(\bar{y}_i)$	S.E. $(\bar{y}_i)$	$V\{S_i(y)\}$	$V(Y_i')$
1-	22	21	0					
6-	26	25	47	1.9				
21-	18	17	191	11.2	.069	± 0.26	47	19,000
51-	26	25	4,051	162.1	.593	± 0.77	192	76,000
151-	20	19	13,899	731.5	5.92	± 2.43	4,004	1,595,000
301-	13	12	23,370	1947.5	34.75	± 5.89	13,898	5,560,000
	125	119			142.3	± 11.93	24,052	10,069,000
							42,193	17,319,000

the variances of the strata means $V(\bar{y}_i)$, and taking the square roots gives the sampling standard errors S.E. $(\bar{y}_i)$ of these means. The means themselves are tabulated in Table 6.5.b.

To obtain the sampling errors of the population estimates we must bear in mind the method by which these estimates were arrived at. The estimate Y of the population total was obtained by multiplying the sample total by the raising factor. We therefore require the variance of the sample total 2011. This will be equal to the sum of the variances of the strata totals. These variances $V\{S_i(y)\}$ are obtained by multiplying s_i^2 by $n_i(1-f)$. The square root of the sum of the variances is then taken and multiplied by 20. Thus

$$\text{S.E.}(Y) = 20 \times \sqrt{42,193} = 20 \times 205.41 = \pm 4110$$

Similarly

$$\text{S.E.}(\bar{y}) = 205.41/125 = \pm 1.64$$

In the case of Y' each $V(\bar{y}_i)$ is multiplied by the square of the number of farms in the size-group, given in Table 6.5.b. Thus

$$142.3 \times 266^2 = 10,069,000$$

Taking the square root of the sum,

$$\text{S.E.}(Y') = \sqrt{17,319,000} = \pm 4160$$

and hence

$$\text{S.E.}(\bar{y}') = 4160/2496 = \pm 1.67$$

It will be noted that although Y' is slightly more accurate than Y the standard error given by the above calculation is slightly greater. There are two reasons for this. In the first place in calculating $\text{S.E.}(Y)$ we have neglected the errors introduced by the use of a working sampling fraction. The average errors from this cause are equivalent to errors introduced by rounding off the numbers in the different strata of the sample to whole numbers. In the second place, use of the exact sampling fractions will result in slightly different contributions to the error variance from the different strata, which may result in raising or lowering the estimate of the standard error.

The computations for number of farms growing wheat follow the same pattern, with the exception that the variance of each size-group total $V(u_i)$ is estimated from the proportion of farms growing wheat in that size-group.

TABLE 7.6.b—ESTIMATION OF THE SAMPLING ERROR OF THE NUMBER OF FARMS GROWING WHEAT FROM EXAMPLE 6.5

Size-group (acres)	n_i	u_i	p_i	$V(u_i)$
1-5	22	0	0	0
6-20	26	1	.03846	.913
21-50	18	5	.27778	3.431
51-150	26	21	.80769	3.837
151-300	20	16	.80000	3.040
301-	13	11	.84615	1.608
				12.829

The calculations for S.E. (U) are given in Table 7.6.b. For the largest size-group, for example,

$$V(u_i) = 13 \times 0.84615 \times 0.15385 \times 19/20 = 1.608$$

We then have

$$\text{S.E. (U)} = 20 \times \sqrt{12.83} = \pm 71.64$$

If the sampling errors of the different strata means are not required, the above calculations can be simplified slightly by omitting the factor $(1-f)$ till the final stage of the computation.

Example 7.6.b

Estimate the sampling errors of the estimate of wheat acreage and number of farms growing wheat obtained by stratification after selection of the random sample of Example 6.6.

The computations follow the same lines as those for S.E. (Y') in Example 7.6.a. The values obtained are :

$$\text{S.E. (Y')} = \pm 4320 \text{ acres}$$

$$\text{S.E. (U')} = \pm 75.2 \text{ farms}$$

The value for S.E. (Y') is slightly greater than that for the similar estimate of Example 7.6.a. The difference, however, is not a precise measure of the relative accuracy of stratification before and after selection. In the first place, since different samples are involved, there are differences in the estimates of the within-strata variances. In particular, the estimate of the variance for size-group 301- is much greater in the random sample because of the one high value 265. These differences are merely errors of estimation and are not a reflection of the relative accuracy of the two samples. On the other hand, the two largest size-groups, which have the highest variances, happen by chance to have more than the proportionate number of farms in the random sample, and this particular random sample will therefore tend to give a more accurate estimate with stratification than will a stratified sample. On the average, however, random samples stratified after selection will give slightly less accurate values than stratified samples.

7.7 Pooled estimate of error : the analysis of variance

In a stratified sample with equal sampling fractions all the f_i are equal. If in addition the within-strata variances σ_i^2 are equal, by putting $\Sigma(n_i) = n$ we obtain the simplified formula

$$\text{S.E. } (\bar{y}) = \sigma_1 \sqrt{\frac{1-f}{n}}$$

This is the same as the formula for a random sample, with the exception that σ is replaced by σ_1 , the common within-strata standard deviation.

The most accurate estimate of σ_1 will be obtained if the estimates of σ_i^2 from the various strata are weighted according to the numbers of degrees of freedom on which they are based. This is equivalent to adding the sums of squares of deviations from the strata means and dividing by the sum of the associated degrees of freedom, *i.e.* by $n - t$.

It is worth noting here an identity which relates the above sum of squares to the sum of squares of deviations from the general mean. This is

$$\sum S_i (y - \bar{y}_i)^2 + \sum n_i (\bar{y}_i - \bar{y})^2 = S (y - \bar{y})^2$$

This is easily verified if we recognize that

$$\sum n_i (\bar{y}_i - \bar{y})^2 = \sum \{ \bar{y}_i S_i (y) \} - \bar{y} S (y)$$

The first term on the right-hand side is the sum of the products of the means and totals of the separate strata, *i.e.* the "corrections for the means" for the separate strata, and the second term is the "correction for the general mean."

The arithmetical computations can be conveniently arranged in the form of what is known as the *analysis of variance*. This is shown schematically in Table 7.7.a.

TABLE 7.7.a—ANALYSIS OF VARIANCE BETWEEN AND WITHIN STRATA

	Degrees of freedom	Sum of squares	Mean square
Between strata . . .	$t - 1$	$\sum \bar{y}_i S_i (y) - \bar{y} S (y)$	A
Within strata . . .	$n - t$	$\sum S_i (y - \bar{y}_i)^2$	$B = s_1^2$
Whole sample . . .	$n - 1$	$S (y^2) - \bar{y} S (y)$	$C = s^2$

The most convenient form of computation, at least when the numbers in the different strata are small, is to calculate the sums of squares for the whole sample and between strata, and obtain the sum of squares within strata by subtraction. The mean squares are then obtained by division by the degrees of freedom. Only the mean square within strata, s_1^2 , is required for the present purpose. The mean square for the whole sample approximates closely to an estimate s^2 of the variance per unit that would result from random sampling of the whole population. The interpretation of the mean square between strata, A , is discussed in Section 8.10.

A further simplification is possible when each of the strata contains only two sampling units. In this case the sum of squares within strata can be calculated directly from the differences of the y 's of the pairs of units within each stratum. If these differences are denoted by d , the sum of squares will be $\frac{1}{2} S (d^2)$, and consequently, since there are t differences each contributing one degree of freedom,

$$s_1^2 = \frac{1}{2} S (d^2) / t$$

The analysis of variance has many applications, not only in sampling but in other fields of statistics. As its name implies, it provides a way of determining

the different components of variance to which a given type of material is subject. As such it is of particular value in investigations of the efficiency of different types of sampling. Its uses in this connection will be explained in Chapter 8.

The above discussion has been based on the assumption that the within-strata variances are equal. In practice this is not likely to be exactly true, though the data at our disposal may be insufficient to determine the true variance law with any accuracy. It is therefore important to ascertain what is the position if a pooled estimate of variance is used when the variances are in fact unequal.

If all the sampling fractions are equal we have, from equation 7.6.b,

$$V(\bar{y}) = \frac{1-f}{n} \frac{\sum (n_i s_i^2)}{n}$$

The second factor is a weighted mean of the s_i^2 , with weights equal to n_i . In the pooled estimate of error described above s_1^2 is a weighted mean of the s_i^2 with weights proportional to $n_i - 1$. Unless the numbers in the different strata are very small, and associated in magnitude with σ_i^2 , there will be little difference between the two estimates. Use of the pooled estimate of error in the estimation of the error of the population mean and total will not, therefore, introduce any serious disturbance when the sampling fractions are equal, even when the within-strata variances are very unequal. On the other hand, the use of a pooled estimate to determine the errors applicable to the mean or total of part of the population, *e.g.* a single stratum mean, may be very misleading, and it is therefore best, when there are marked differences in the within-strata variances, and the numbers in the different strata are not too small, to keep the error estimates separate, as has been done in Example 7.6.a.

There is, of course, nothing sacrosanct in weighting by the degrees of freedom; these weights merely give the most accurate estimate when the variances are equal, and enable the analysis of variance technique to be used. If the n_i are too small for separate estimates of the within-strata variances to be of value, we can still use a pooled estimate, weighting by n_i if this appears advisable.

The situation is completely different with a variable sampling fraction. In this case equation 7.6.a shows that weights proportional to $g_i^2 n_i (1 - f_i)$ are required. The pooled estimate of variance may therefore be decidedly misleading, even with quite small differences in the within-strata variances. Consequently in this case separate estimates with proper weighting should always be used.

Example 7.7.a

Estimate the sampling errors of the estimate of the total wheat acreage and number of farms growing wheat from the stratified systematic sample with a variable sampling fraction of Example 6.7.

Since the systematic selection was from a list arranged by districts the sample will substantially be stratified by districts as well as size-groups. The numbers in the 6-20 and 21-50 size-groups are too small for the district stratification to be effective, but for the larger size-groups the within-districts variance is required, instead of the overall variance within the size-group.

The available number of degrees of freedom in each district and size-group is small, and inspection of the values of Table 6.7.a shows that there are no marked differences in variability in the different districts. We may therefore appropriately use a pooled estimate of variance for each size-group.

The district totals and means for the largest size-group are shown in Table 7.7.b.

TABLE 7.7.b—DISTRICT TOTALS AND MEANS FOR SIZE-GROUP 501—

District :	1	2	3	4	5	6	7	All
No. of farms	1	4	2	9	0	1	0	17
Total .	114	487	315	1937	—	72	—	2925
Mean .	114	121.75	157.5	215.2222	—	72	—	172.0588

The analysis of variance is given in Table 7.7.c. The sum of squares between districts is $114 \times 114 + 487 \times 121.75 + \dots - 2925 \times 172.0588 = 40,698$. The total sum of squares is $114^2 + 119^2 + 107^2 + \dots - 2925 \times 172.0588 = 72,067$. Subtraction gives the within-district sum of squares, and division by the degrees of freedom the mean squares.

TABLE 7.7.c—ANALYSIS OF VARIANCE BETWEEN AND WITHIN DISTRICTS OF THE WHEAT ACREAGES OF SIZE-GROUP 501— (DATA OF TABLE 6.7.a)

	Degrees of freedom	Sum of squares	Mean square
Between districts .	4	40,698	10,174
Within districts .	12	31,369	2,614
Whole size-group .	16	72,067	4,504

The within-district mean square is substantially less than the overall mean square, indicating the greater similarity of farms within a district and consequent gain in accuracy by stratification.

The size-groups 51-, 151-, and 301- can be analysed in the same manner. There is some difference in mean squares for size-group 301-, but little difference for the other two size-groups. For size-groups 6- and 21- the overall variability within size-groups can be taken.

The remainder of the computations are set out in Table 7.7.d. They follow the same lines as Example 7.6.a, with the exception that the factors $(1 - f_i)$ are different, and that the variance of each group total must be multiplied by the square of the raising factor for that group, and the resultant

TABLE 7.7.d—ESTIMATION OF SAMPLING ERRORS OF THE ESTIMATES OF EXAMPLE 6.7

Size-group (acres)	n_i	s_i^2	$V\{S_i(y)\}$	g_i^2	$V(Y_i)$
1-5	0				
6-20	3	0	0	40,000	0
21-50	6	53.5	320	3,600	1,152,000
51-150	26	159.2	3,930	400	1,572,000
151-300	40	564.2	20,310	100	2,031,000
301-500	43	1,703	58,580	25	1,464,000
501-	17	2,614	29,630	9	267,000
	135				6,486,000

variances added, in accordance with formula 7.5.a. The estimated standard error of the total acreage is thus $\sqrt{6,486,000} = \pm 2550$.

It will be noted that the acreage of wheat in the smallest size-group has been assumed to be zero, and that the estimated zero error variance of the second size-group is based on only two degrees of freedom, and is therefore very inaccurately determined. It is clear, however, from the nature of the material and the trend of the variances in the larger size-groups that this variance must be small.

In the computation of the standard error of the number of farms growing wheat, allowance should also strictly be made for the stratification by districts. If the number of farms in each size-group district sub-class were large this could be done by calculating the variance of each size-group total of farms growing wheat by the method of Example 7.6.a. The numbers in many of the sub-classes are so small, however, that the approximation resulting from using the estimated proportions p to calculate the variances will be unsatisfactory. In this case it will be sufficient to ignore the district stratification, calculating the variance of each size-group total of farms growing wheat from the proportion in that size-group, and then proceeding in the same manner as for wheat acreage. The resultant standard error will be found to be ± 88.9 .

Example 7.7.b

Estimate the sampling standard errors of the regional and varietal means of the yields of potatoes given in Table 5.23.a, and compare them with the standard errors already obtained in Example 7.5.

As mentioned in Section 5.23, the sample can be regarded as stratified by regions (but not by varieties). The regional standard errors are therefore derived from the analysis of variance within and between regions. This is

given by lines (1), (4) and (5) of Table 7.7.e. The required standard errors are therefore $\sqrt{(5.25/174)} = \pm 0.17$, etc.

TABLE 7.7.e—POTATO SURVEY: ANALYSIS OF VARIANCE OF YIELDS PER ACRE

	Degrees of freedom	Sum of squares	Mean square
Between regions (1)	4	173.7	43.42
Within regions { Between varieties (2) .	16	987.7	61.73
{ Within varieties (3) .	880	3713.6	4.22
{ Total (4)	896	4701.3	5.25
Total (5)	900	4875.0	5.42
Between varieties (6)	4	887.3	221.82
Within varieties { Between regions (7) .	16	274.1	17.13
{ Within regions (8) .	880	3713.6	4.22
{ Total (9)	896	3987.7	4.45
Total (10)	900	4875.0	5.42

The mean square within regions, 5.25, is 1.24 times the mean square within regions and varieties, 4.22, already given in Example 7.5. This latter mean square is obtained from an analysis of variance within and between the regional-variety groups, lines (2) and (3). This would have been the appropriate mean square for estimating the errors of the regional means if the sample had been stratified by regions and varieties.

The exact standard errors of the varietal means cannot be obtained by any simple process. If the sample were fully random, and not stratified by regions, the correct estimate would be that given by the within-varieties component of variance, *i.e.* by treating the sample as if it were stratified by varieties. Stratification by regions will reduce the sampling error of the varietal means slightly, but not to any great extent. Consequently the estimate obtained by stratifying by varieties and not by regions will be somewhat of an overestimate of the true standard error.

The analysis of variance within and between varieties (ignoring regions) is given by lines (6), (9) and (10) of Table 7.7.e. Approximations to the varietal standard errors are therefore given by $\sqrt{(4.45/393)} = \pm 0.11$, etc.

It should be noted that although the sample is stratified by regions the component of variance due to regions must *not* be eliminated when calculating the standard errors of the varietal means. This must only be done if the sample is stratified by both varieties and regions. The reason is as follows. If the sample is stratified by both varieties and regions the proportions of fields from each region in each varietal mean will be exactly equal to the proportions for

that variety in the country. Hence only variation between fields of the same variety within each region contributes to the error. In the present case, however, the proportions of fields from each region in each varietal mean do not correspond exactly to the proportions in the country. The deviations are in fact only slightly less than would be obtained in a random sample.

Those familiar with the use of the analysis of variance in replicated experiments may wonder why two separate analyses are required, instead of the single analysis analogous to the partition of the degrees of freedom of a complete 5×5 table into

Regions	4
Varieties	4
Regions $\times$ varieties	16

The reason is that regions and varieties are not *orthogonal*, owing to the differing numbers of fields in the different sub-classes. The analysis of variance of non-orthogonal material is inherently more complicated, and in particular the *interaction* component, regions $\times$ varieties, can only be obtained by rather elaborate calculation (Yates, 1934, A). It is *not* given by the subtraction of the sums of squares for regions (1) and varieties (6) from the sum of squares for all regional-variety sub-classes, (1) and (2), or (6) and (7).

All the components of variance due to the regional-variety classification must be eliminated when calculating the sampling standard errors of varietal differences freed from regional effects, or of regional differences freed from varietal effects, since we are then concerned only with the component of variance within sub-classes. These standard errors have already been discussed in Example 7.5. The within-sub-classes component can be obtained by splitting the sum of squares within each region into between and within varieties and pooling the components so obtained, as in lines (2) and (3) of Table 7.7.e; by doing the same for regions within varieties, as in lines (7) and (8); or by making a direct analysis between and within regional-variety sub-classes. All three processes are equivalent arithmetically, and give the same sum of squares (880 degrees of freedom) within sub-classes.

The reader should calculate for himself the various sums of squares (other than the total sum of squares) given in Table 7.7.e. These can be obtained from the data of Table 5.23.b. For this purpose the means require to be recalculated to a greater number of decimal places than those given in Table 5.23.b.*

It will now be seen that three separate variances are relevant to the calculation of the standard errors appropriate to the various estimates we have obtained from the data of Table 5.23.b. There is also a fundamental difference in the nature of the errors. Those appropriate to the regional means and varietal means are genuine sampling errors. They answer the question:

* There are one or two minor last-place discrepancies between the means and totals given in Table 5.23.b. These are due to reconstruction of the data from a table in a report.

given the existing distribution of varieties in the different regions, by how much may the sample regional (or varietal) means be expected to deviate from the corresponding means over all fields? If the sampling fraction were large the correction for finite sampling would require to be made.

In evaluating the errors of the estimates of the regional and varietal differences freed from the effects of the other factor, we are concerned with the more general question: how far are the estimated differences likely to be in error due to chance variations between the fields on which the varieties are grown? This is not a sampling error. It would still exist if the data collected represented all the potato fields in the country. The correction for finite sampling should therefore not be applied.

It should be noted also that this latter estimate of error is based on the assumption that the distribution of the varieties over the different fields within a region is random, and that conditions of growth, etc., are also randomly distributed. This may be far from the truth. If variety P is regarded, rightly or wrongly, as particularly suitable for poor soils, it will tend to be grown on poor soils, and its yield will be less for this reason. A new variety will tend to be grown by the more progressive farmers, and may in consequence give higher yields, even though no better than the older varieties. Consequently the estimates of error merely provide lower limits to the real errors; in other words they represent the errors attributable to the residual random component of variance only. Consequently, as already emphasized in Section 5.23, all conclusions must be tentative. Only experiments can give definite answers.

7.8 Ratio method : random sample

In order to calculate the variance of $\bar{r}$ the correlation between the values of x and y for the same sampling unit must be taken into account. From formula 7.5.k, with the additional covariance term and allowance for finite population, we obtain for a random sample

$$V(\bar{r}) = \frac{1-f}{n} \bar{r}^2 \left(\frac{V(y)}{\bar{y}^2} - \frac{2 \text{cov}(xy)}{\bar{x}\bar{y}} + \frac{V(x)}{\bar{x}^2} \right)$$

This is an approximate formula, but is accurate enough for practical purposes in the cases met with in sampling.

The formula can be put in an alternative form, which somewhat simplifies the approach to more complicated cases such as stratified samples. If we denote by Q the sum of the squares of the deviations of the y 's from the values given by the ratio line (OMD of Fig. 6.8), we have

$$\begin{aligned} Q &= S(y - \bar{r}x)^2 \\ &= S\{(y - \bar{y}) - \bar{r}(x - \bar{x})\}^2 \\ &= S(y^2) - 2\bar{r}S(xy) + \bar{r}^2S(x^2) \\ &= S(y - \bar{y})^2 - 2\bar{r}S(x - \bar{x})(y - \bar{y}) + \bar{r}^2S(x - \bar{x})^2 \end{aligned}$$

the last two expressions being those which are suitable for computation.

If we now take s_q^2 to represent the estimated mean square deviation from the true ratio line, we have

$$s_q^2 = Q/(n-1)$$

We then find

$$V(\bar{r}) = \frac{1-f}{n\bar{x}^2} s_q^2 = \frac{1-f}{\{S(x)\}^2} ns_q^2$$

$$V(Y) = \frac{\bar{X}^2}{\{S(x)\}^2} (1-f) ns_q^2$$

$$V(\bar{y}) = \frac{\bar{x}^2}{\bar{x}^2} \cdot \frac{1-f}{n} s_q^2$$

The first of the above formulæ is equivalent to formula 7.5.g.

The analogy with the case of a random sample without supplementary information can now be seen. Apart from the factor $\bar{x}^2/\bar{x}^2$, which will be approximately unity, the only difference is that the sum of the squares of the deviations of y from the mean $\bar{y}$ of the sample is replaced by the sum of the squares of the deviations of y from the ratio line of the sample.

The variance of a standardized estimate y_0 will be obtained by replacing x by x_0 in the above formula.

In the case of two-phase sampling the first-phase estimate $\bar{x}_1$ of $\bar{x}$ will be subject to sampling errors. If the sampling is random for both phases the variance of $\bar{y}$ will be

$$V(\bar{y}) = \frac{1}{n_2} \left(1 - \frac{n_2}{n_1}\right) \frac{\bar{x}_1^2}{\bar{x}_2^2} s_q^2 + \frac{1-f_1}{n_1} s^2$$

where the suffices refer to the phases. s_q^2 is calculated as above from the second-phase units, and s^2 is the total variance of y , also calculated from the second-phase units.

The first term of the above formula will be recognized as the sampling variance of $\bar{y}$ due to the second-phase sampling of the first-phase sample (regarded as without error), while the second term is the first-phase sampling variance of $\bar{y}$, i.e. the variance which would be obtained if y were determined for all units of the first-phase sample. This subdivision of the variance provides a general method of obtaining the errors of two-phase sampling. In certain types of sampling the circumstance that values of y are not available for all the first-phase units introduces complications into the estimation of the second component of variance which will be dealt with in Section 8.7.

It frequently happens that the available supplementary information is out of date or otherwise subject to error. If, however, values of x are known for all units of the population these values can be used in the calculation of $\bar{x}$ from the selected units. If this is done bias will be avoided, and the effect of these errors on the final estimates will be correctly assessed, provided the original frame is complete. If the frame is not complete the sampling divides into two parts: that covering units included in the original frame, for which the ratio or regression method of estimation can be used; and that covering units not

so included, for which the appropriate method of estimation without supplementary information will be required.

It may also be noted that if the variance $V_x(r)$ of r for fixed x is constant over the whole range of values of x , and if r itself exhibits no trend over this range, $V(\bar{r})$ may be estimated from formulæ 7.5.h and 7.5.e, substituting x for w and r for y . This method of estimation has the advantage of saving computation in cases in which the values of r are directly available while those of y are not, but it will give a biased estimate of error if the above conditions are not fulfilled. An example of the method is given for a stratified sample in Example 7.17.

If $V_x(r)$ is virtually constant, the sampling error of any ratio estimate can be rapidly calculated once the value of $V_x(r)$ has been established, since only $S(x)$ and $S(x^2)$ require to be known, formula 7.5.e being used. Similarly, if $V_x(r)$ is inversely proportional to x , i.e. equal to λ/x , formula 7.5.f can be used, only $S(x)$ being required. The effective constancy of λ , and its value, can be most simply established by calculating $V(\bar{r})$ in the ordinary manner for various batches of data and calculating the resultant values of λ from formula 7.5.f. This is in general preferable to using formulæ 7.5.i and 7.5.f directly.

Example 7.8.a

Estimate the sampling error of the ratio estimate of the acreage of wheat from the random sample of farms (Example 6.9.a).

We have

$$\begin{array}{lll} S(y^2) = 207,261 & S(xy) = 902,958 & S(x^2) = 5,061,734 \\ \bar{r} = .1522430 & 2\bar{r} = .3044860 & \bar{r}^2 = .0231779 \\ Q = 49,643 & s_q^2 = 400.35 \end{array}$$

$$\text{S.E.}(\bar{r}) = \frac{1}{15,114} \sqrt{\{(1 - 1/20) 125 \times 400.35\}} = \pm 0.01443$$

$$\text{S.E.}(Y) = 273,074 \times 0.01443 = \pm 3,940$$

Example 7.8.b

Estimate the sampling error of the estimates of Example 6.9.b.

We have

$$\begin{array}{llll} n = 43 & f = 43/325 = 0.1323 & g = 7.5581 & \\ S(y) = 799 & \bar{y} = 18.5814 & \bar{x} = 79.6977 & S(x) = 3,427 \\ S(y^2) = 22,065 & & S(xy) = 76,965 & S(x^2) = 328,323 \\ \bar{y} S(y) = 14,846.5 & & \bar{y} S(x) = 63,678.5 & \bar{x} S(y) = 273,124.0 \\ S(y - \bar{y})^2 = 7218.5 & S(x - \bar{x})(y - \bar{y}) = 13,286.5 & S(x - \bar{x})^2 = 55,199.0 & \\ \bar{r} = 0.233149 & 2\bar{r} = 0.466298 & \bar{r}^2 = 0.0543585 & \\ Q = 4023.6 & s_q^2 = 95.80 & & \end{array}$$

$$\text{S.E.}(100\bar{r}) = \frac{100}{3427} \sqrt{(0.8677 \times 43 \times 95.80)} = \pm 1.744$$

Had the sample been a random sample of individuals the formula of Section 7.4 would have been applicable, giving

$$\text{S.E. } (100 \bar{r}) = 100 \sqrt{(0.8677 \times 0.2331 \times 0.7669/3427)} = \pm 0.673$$

The large difference between these two standard errors is an indication of the additional variability between kraals, and illustrates the misleading results that may be obtained by using the formula of Section 7.4 when the sampling units consist of groups of individuals and not single individuals.

Since the total number of persons in the reserve is unknown, the standard errors of the total numbers are derived from the formulæ appropriate to a random sample without supplementary information. We therefore have

$$\begin{aligned} S(y - \bar{y})^2/(n-1) &= 171.87 & S(x - \bar{x})^2/(n-1) &= 1,314.3 \\ \text{S.E. } (X) &= 7.5581 \times \sqrt{(0.8677 \times 43 \times 1314.3)} = \pm 1,673 \\ \text{S.E. } (Y) &= 7.5581 \times \sqrt{(0.8677 \times 43 \times 171.87)} = \pm 605.3 \end{aligned}$$

The standard error of the number present in the reserve can be calculated in the same manner from the sum of the squares of the deviations of $(x - y)$, which in turn can be calculated directly from the separate values of $(x - y)$. In the present case, where the separate values of $(x - y)$ are not tabulated, and where $S(x - \bar{x})(y - \bar{y})$ has already been calculated, it is more convenient to obtain the required sum of squares of deviations from the sums of squares and products already calculated (see Section 7.5). Thus

$$\begin{aligned} S\{(x - y) - (\bar{x} - \bar{y})\}^2 &= 55,199.0 + 7,218.5 - 2 \times 13,286.5 = 35,844.5 \\ \text{S.E. } (X - Y) &= 7.5581 \times \sqrt{\{0.8677 \times 43 \times 35,844.5/42\}} = \pm 1349 \end{aligned}$$

Note that x and y are not independent, being derived from the same sampling units, and therefore $V(X - Y)$ is not equal to $V(X) + V(Y)$, but to $V(X) + V(Y) - 2 \text{ cov}(XY)$. Putting $\text{cov}(XY)$ equal to $13,286.5/42$ gives the same result as above.

7.9 Ratio method : stratified sample with uniform sampling fraction

(a) When the ratio is assumed to be the same for all strata :

Instead of taking the sum of squares of deviations from the general ratio line, the deviations from a series of lines parallel to this line and passing through the points representing the strata means must be taken, the divisor $n - 1$ being replaced by $n - t$.

Thus

$$\begin{aligned} Q &= \sum S_i \{(y - \bar{y}_i) - \bar{r}(x - \bar{x}_i)\}^2 \\ &= \sum S_i (y - \bar{y}_i)^2 - 2\bar{r} \sum S_i (y - \bar{y}_i)(x - \bar{x}_i) + \bar{r}^2 \sum S_i (x - \bar{x}_i)^2 \\ s_q^2 &= Q/(n - t) \end{aligned}$$

The sums of squares and products will be recognized as the sums of squares and products within strata, similar to those already obtained for the y variate

in the pooled estimate of error for a stratified sample without supplementary information.

(b) When the ratio is permitted to assume different values for the different strata :

The common $\bar{r}$ is replaced by $\bar{r}_i$ corresponding to the different strata, so that

$$Q = \sum S_i (y - \bar{y}_i)^2 - 2 \sum \bar{r}_i S_i (y - \bar{y}_i) (x - \bar{x}_i) + \sum \bar{r}_i^2 S_i (x - \bar{x}_i)^2$$

The divisor $n - t$ stands.

In this case the contribution to Q from each stratum is best computed separately. If desired the variances of the contributions to Y from the different strata may also be estimated separately. This course is equivalent to assigning slightly different weights to the different contributions to Q , the situation being analogous to that already discussed in Section 7.6.

Note that if the population totals X_i are not known for the different strata but the total X for the whole population is known, the formulæ for case (a) must be used for calculating the sampling errors of $\bar{r}$ and Y , even if the ratio clearly varies from stratum to stratum, since the method of estimation must be that corresponding to case (a).

Example 7.9

Estimate the sampling errors of the estimate of Example 6.10.

The contributions to Q from the six districts are :

District	Q_i	District	Q_i
1 ..	5,107.59	4 ..	20,566.56
2 ..	1,550.71	5 & 7 ..	3,737.14
3 ..	7,963.98	6 ..	1,080.92
			TOTAL 40,006.90

Hence

$$\text{S.E. } (Y) = \frac{273,074}{15,114} \sqrt{\left\{ \left(1 - \frac{1}{20}\right) 125 \frac{40,006.90}{119} \right\}} = 3,610$$

7.10 Ratio method : stratified sample with variable sampling fraction

If the ratio is assigned different values in the different strata the variance of the estimated total is

$$\begin{aligned} V(Y) &= \sum \left[\frac{X_i^2}{\{S_i(x)\}^2} (1 - f_i) n_i s_{qi}^2 \right] \\ &= \sum \{g_i^2 (1 - f_i) n_i s_{qi}^2\} \text{ approximately,} \\ V(\bar{r}) &= V(Y)/X^2 \end{aligned}$$

Here the s_{qi}^2 are estimated separately for each stratum, using the value of the ratio appropriate to the stratum and the divisor $n_i - 1$ for Q_i .

If the ratio is assumed to be the same for all strata the same formula may be used, with the exception that the Q_i are calculated using the general ratio, with divisors $n_i - 1$ as before.

For an illustration of the application of these formulæ see Example 7.17.

7.11 Ratio method : integral values of the supplementary variate

When the supplementary variate x can only assume small integral values the above formulæ for Q can be simplified by classifying the data according to the value of x . The most common instance in censuses and surveys is in surveys of human populations in which the sampling units are households and information is required on individuals.

In the analysis of data appertaining to individuals the working unit in the analysis will commonly be the individual, although the sampling unit is the household. Clear distinction must therefore be made between the values y for the sampling units which in households of two, for instance, will consist of the totals of pairs of individuals, and the values for the individuals. These latter values we will denote by z , with the convention that $[z]$ for families of more than one unit represents the total of the individuals in this family, so that $[z]$ equals y . With this notation $\bar{r} = \bar{z}$. Suffices will be used to indicate size of family ; $n_1, n_2, \dots$ to denote the numbers of *families* of the different sizes.

No difficulty should be found in transforming the formulæ for Q into a form suitable for computation. In the case of a random sample, for instance, we find

$$\begin{aligned} Q &= S_1(y^2) + S_2(y^2) + \dots - 2\bar{r} \{S_1(y) + 2S_2(y) + \dots\} \\ &\quad + \bar{r}^2(n_1 + 4n_2 + \dots) \\ &= S_1(z^2) + S_2[z]^2 + \dots - 2\bar{z} \{S_1(z) + 2S_2(z) + \dots\} \\ &\quad + \bar{z}^2(n_1 + 4n_2 + \dots) \\ &= S_1(z - \bar{z}_1)^2 + S_2([z] - 2\bar{z}_2)^2 + \dots + n_1(\bar{z}_1 - \bar{z})^2 \\ &\quad + 4n_2(\bar{z}_2 - \bar{z})^2 + \dots \end{aligned}$$

It will be noted that in order to calculate Q the quantities $[z]$ are required. In the event of the survey material being recorded on punched cards, each individual will normally be assigned to a separate card. The required totals can then be obtained on the tabulator by sorting for family designation, family size, and any stratification which is required, and controlling on family designation, either printing the totals or reproducing them directly on to new family cards. The latter procedure will be advantageous if further analysis is required on the characteristics of families regarded as entities.

This type of analysis can be confined to a special type of individual, such as adult males. If punched cards are used a card count will have to be introduced to count the number of the special type occurring in each family,

which may be termed the "partial size" of the family. There is then the minor complication that families of different partial sizes will occur together in the tabulated results. This is overcome if the results are reproduced on cards, as the "partial sizes" can be punched on the cards, and the cards subsequently sorted by partial size and listed.

The last of the above forms for Q separates the various components of variance. The first term gives the contribution to Q arising from variation between families of one, the second that between families of two, etc., and the first term of the second set gives the contribution due to the average deviation of families of one from the general mean, etc. If the sample were stratified for size of family the second set of terms would be omitted. They will also be omitted if the error of a mean standardized for distribution of family size is required.

The above formula for Q can be set out in analysis of variance form, as in Table 7.11.

TABLE 7.11—ANALYSIS OF VARIANCE FORM FOR Q FOR INTEGRAL VALUES OF x

	Degrees of freedom	Sum of squares
Between families of size 1	$n_1 - 1$	$S_1(z - \bar{z}_1)^2$
„ „ „ „ 2	$n_2 - 1$	$S_2([z] - 2\bar{z}_2)^2$
„ „ „ „ 3	$n_3 - 1$	$S_3([z] - 3\bar{z}_3)^2$
Between means of families of different sizes	$t - 1$	$n_1(\bar{z}_1 - \bar{z})^2 + 4n_2(\bar{z}_2 - \bar{z})^2 + \dots$

The mean square of the first line then gives the estimate s_1^2 of the error variance of families of size 1, the mean square of the second line, divided by 4, the estimate of the error variance of family means of size 2, etc. The means of families of a given size can thus be compared for different parts of the population, remembering that the further divisors for s_2^2 , etc., depend on the numbers of families and not individuals entering into the means.

7.12 Regression method : random sample

The estimation of error in the regression method follows much the same lines as in the ratio method. The sum of squares of deviations from the ratio line is replaced by the sum of squares of deviations from the regression line, and the divisor $n - 2$ is used instead of $n - 1$, since an additional degree of freedom is accounted for by the fact that the regression line not only passes through the mean point, but has its slope determined independently from the data.

The sum of the squares of the deviations from the regression line is given by the equation

$$\begin{aligned} Q &= S(y - y_i)^2 \\ &= S(y - \bar{y})^2 - b S(x - \bar{x})(y - \bar{y}) \\ &= S(y - \bar{y})^2 - \frac{\{S(x - \bar{x})(y - \bar{y})\}^2}{S(x - \bar{x})^2} \end{aligned}$$

so that

$$s_i^2 = Q/(n - 2)$$

and, if errors in b are neglected

$$V(\bar{y}) = \frac{1 - f}{n} s_i^2$$

The error variance of b , if the variance of y for fixed x is constant, is

$$V(b) = \frac{s_i^2}{S(x - \bar{x})^2}$$

so that if the regression is truly linear the error variance of a standardized value y_0 of y for the value x_0 of x is

$$V(y_0) = \left\{ \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S(x - \bar{x})^2} \right\} s_i^2$$

The correction for finite population is here omitted since standardized values are ordinarily used for comparative purposes.

Allowance for errors in b can be made in $V(\bar{y})$ in the same way, but such errors will always be small relative to the other component of error. They will on the average increase the error variance approximately in the ratio $n/(n - 1)$.

In the case of two-phase sampling the sampling variance of $\bar{x}_1$ will introduce the additional term $b^2 V(\bar{x}_1)$ into the above expression for $V(\bar{y})$. The general approach is given in Section 8.7.

If an arbitrary value b_0 of the regression coefficient is used, instead of the value b calculated from the data, the formula for the sum of the squares of the deviations from the arbitrary regression line becomes

$$Q = S(y - \bar{y})^2 - 2b_0 S(x - \bar{x})(y - \bar{y}) + b_0^2 S(x - \bar{x})^2$$

and

$$s_i^2 = Q/(n - 1)$$

This procedure is equivalent to the analysis of the values $y - b_0 x$ from the individual sampling units. Use of the above expression for Q saves the trouble of calculating the values of $y - b_0 x$ for the individual sampling units, at the expense of calculating the sums of squares and products of x and y instead of the sums of squares of $y - b_0 x$.

No allowance has to be made for errors in b_0 , but an arbitrary value b_0 should not be used for standardization unless it is known that b_0 approximates closely to b , so that the error $(b_0 - b)(x_0 - \bar{x})$ introduced into the standardization correction is small.

Example 7.12.a

Estimate the errors of the estimates of Example 6.12.a.

We have

$$\begin{aligned} Q &= 164,904 - 0.19316 \times 624,739 = 44,229 \\ s_t^2 &= Q/123 = 359.59 \\ V(\bar{y}) &= (1 - 1/20) 359.59 / 125 = 2.733 = (\pm 1.653)^2 \\ \text{S.E.}(Y) &= 2496 \times 1.653 = \pm 4126 \\ V(b) &= \frac{359.59}{3,234,270} = 0.0001112 = (\pm 0.01054)^2 \end{aligned}$$

Example 7.12.b

Estimate the error of the estimates of total volume of timber of Example 6.12.b.

Except for the estimate derived from the arbitrary value of the regression coefficient $b_0 = 1$ the computations follow the lines already given and are left as an exercise for the reader.

When $b_0 = 1$ we have

$$\begin{aligned} Q &= 115,266 - 2 \times 52,069 + 82,296 = 93,424 \\ s_t^2 &= 93,424/24 = 3893 \end{aligned}$$

The values of the error variance per unit, and the resultant standard errors of the various estimates, are as follows:

	Variance per unit	S.E. (total volume)	Relative efficiency
Sample plots only	4,803	$\pm 710,000$ cu. ft.	72
Ratio method	4,230	$\pm 602,000$ cu. ft.	82
Regression, $b_0 = 1$	3,893	$\pm 639,000$ cu. ft.	89
Regression, $b = .6327$	3,579	$\pm 613,000$ cu. ft.	96
Regression, $b_0 = .55$	3,454	$\pm 603,000$ cu. ft.	100

The relative efficiency of the various methods of estimation is inversely proportional to the value of the variance per unit. Setting the last value at 100 we obtain the relative efficiencies of the last column. The relative efficiencies fall in the order given. If the information from the sample plots only is used, neglecting the eye estimates, about 40 per cent. more sample plots will be required to give results of the same accuracy as those obtained by using a regression of 0.55 on the eye estimates.

The values for $b_0 = 0.55$ have been included to illustrate the fact that any value of the regression coefficient near to the value derived from the data will give results which are of about the same accuracy. Here there is an

apparent small gain in accuracy owing to change in the degrees of freedom from 23 to 24. There is therefore no point in attempting to take account of small differences in the regression coefficient for different parts of the data, or to determine b very exactly. Any value reasonably near the correct value will give a satisfactory adjustment.

The ratio method has given an estimate of the standard error which is relatively low because $S(x)$ for the sample is high. The average performance of the ratio method may best be judged by the variance per unit.

Limits of error can be assigned to the possible bias in the eye estimates by calculating the standard error of the difference $\bar{x} - \bar{y}$, or of the ratio $\bar{r}$. We have

$$\text{S.E.}(\bar{x} - \bar{y}) = \sqrt{(3893/25)} = \pm 12.5.$$

The actual difference, -15.3 , is 1.22 times its standard error, and these data therefore do not by themselves furnish conclusive evidence of the existence of bias in the eye estimates. Taking limits of error of $\pm$ twice the standard error gives limits to the bias of -40.3 and $+9.7$, i.e. -27 per cent. and $+7$ per cent.

Similarly $\text{S.E.}(\bar{r}) = \pm 0.098$, and since $\bar{r} = 1.116$ the ratio of the deviation of $\bar{r}$ from unity to its standard error is 1.19, which compares with the value of 1.22 for $\bar{x} - \bar{y}$.

As mentioned in Example 6.12.b, the more extensive data of the full survey confirmed the existence of bias, giving at the same time a more accurate determination of its average magnitude and variation for different types of woodland.

7.13 Regression method : stratified and balanced samples

- (a) Uniform sampling fraction, regression coefficient the same for all strata :

As for a random sample, except that

$$\begin{aligned} Q &= \sum S_i (y - \bar{y}_i)^2 - b \sum S_i (y - \bar{y}_i) (x - \bar{x}_i) \\ s_i^2 &= Q/(n - t - 1) \\ V(b) &= s_i^2 / \sum S_i (x - \bar{x}_i)^2 \end{aligned}$$

If an arbitrary value b_0 is taken, the formula for Q must be rewritten in the same manner as in an unstratified sample, the divisor being $n - t$.

- (b) Uniform sampling fraction, regression coefficients different for the different strata :

$$b_i = S_i (y - \bar{y}_i) (x - \bar{x}_i) / S_i (x - \bar{x}_i)^2$$

If a pooled estimate of error is used,

$$\begin{aligned} Q_t &= \sum S_i (y - \bar{y}_i)^2 - \sum b_i S_i (y - \bar{y}_i) (x - \bar{x}_i) \\ s_i^2 &= Q/(n - 2t) \\ V(b_i) &= s_i^2 / S_i (x - \bar{x}_i)^2 \end{aligned}$$

If the variances from the regression lines in the different strata are likely to be different, separate estimates of σ_l^2 should be used when evaluating $V(b_i)$.

(c) Variable sampling fraction :

The method is similar to that outlined for the ratio method.

(d) Balanced sample :

As far as the estimation of error is concerned, a balanced sample must be treated as if it were an unbalanced sample from which estimates have been derived by the use of regression on the balanced variate. There will be no regression adjustment to the estimates of the population values, since $\bar{x} - \bar{x}$ will be zero because of the balancing.

7.14 Calibration of eye estimates

When a regression is used to calibrate eye estimates, as described in Section 6.15, the sampling variance of $\bar{y}$ can be split into three parts, that due to errors in b' , that due to the sampling variance of $\bar{x}_1$ arising from the main sampling process, and that due to the variance of $\bar{x}_1 - \bar{x}$ arising from the variance about the regression line.

The component of variance due to errors in b' is usually sufficiently small to be neglected. To a first approximation it equals

$$(\bar{x}_1 - \bar{x})^2 V(b')/b'^4$$

where $V(b')$ is calculated in the ordinary manner from the regression.

The variance of $\bar{x}_1$ is calculated from the values of x for all the selected units in the manner appropriate to the method of sampling adopted. The contribution to $V(\bar{y})$ from this source is approximately

$$V(\bar{x}_1)/b'^2$$

A closer approximation is obtained by multiplying this variance by $\{V(x) - V_l(x)\}/V(x)$, where $V(x)$ is that part of the variance of x which contributes to $V(\bar{x}_1)$ and $V_l(x)$ is the residual variance of x about the regression line.

The variance of $\bar{x}_1 - \bar{x}$ due to variance about the regression line is calculated from the residual variance $V_l(x)$ of x about this line. If n_1 and n represent the numbers of units in the original sample and the sub-sample for eye estimates, the contribution to $V(\bar{y})$ when all units are given equal weight in the mean is

$$(n_1 - n) V_l(x)/b'^2 nm_1$$

If the x 's are weighted according to area a or other weights, the last expression becomes

$$\frac{V_l(x)}{b'^2} \left(\frac{S'(a)}{S_1(a)} \right)^2 \left[\frac{S(a^2)}{\{S(a)\}^2} + \frac{S'(a^2)}{\{S'(a)\}^2} \right]$$

where S_1 , S and S' indicate summation over the whole sample, over the

sub-sample, and over the part of the sample not included in the sub-sample, respectively.

Example 7.14

Estimate the variance of the mean yield per acre obtained in Example 6.15.

The component of variance due to errors in b' is negligible, since $\bar{x}_1 - \bar{x}$ is nearly zero.

Since $\bar{x}_1$ was calculated by weighting the weighted mean eye estimates of the individual farms by the wheat acreages of these farms, the ratio method of calculating the sampling error at the first stage is applicable. A table was therefore prepared giving for each farm, (1) the total wheat acreage, (2) the weighted mean yield per acre based on the eye estimates of all the chosen fields on that farm, and (3) the product of these two numbers. Columns (1) and (3) constitute, in the ordinary ratio notation, the x and y values. Using these tabulated values, and the ordinary formula for the variance of a ratio, with the inclusion of the factor $(1-f)$, we find $V(\bar{x}_1) = 0.8200$,* and consequently the corresponding component of variance is $0.8200/0.6926^2$ or 1.7095 . The factor $(1-f)$ can properly be included here since for the majority of farms all the fields were taken. On the other hand, although the variance per field of the eye estimates is probably reasonably constant, the alternative approach outlined in Section 7.5, making use of this fact, would present difficulties, since the sampling units at the first stage are farms and not fields. The direct approach is therefore simpler.

The residual variance about the regression line was found to be, from an analysis of the unweighted data for fields, $V_l(x) = 7.038$. The sums and sums of squares of the areas of the individual fields for which actual yields are and are not available, and of all fields, are

$$\begin{array}{lll} S(a) = 610 & S'(a) = 1279 & S_1(a) = 1889 \\ S(a^2) = 15,172 & S'(a^2) = 33,899 & S_1(a^2) = 49,071 \end{array}$$

Substitution in the formula above gives a component of variance of 0.4137 . Fields and not farms can reasonably be used here, since errors in the eye estimates may be expected to be reasonably independent from field to field. This is not the case with $V(\bar{x}_1)$, since the yields of fields on the same farm often show considerable correlation.

The standard error of the adjusted mean yield is therefore $\sqrt{1.7095 + 0.4137} = \pm 1.46$ bushels per acre. The main source of error is that due to sampling errors introduced by the variation in yields from farm to farm. The eye estimates are shown to be sufficiently consistent and to give adequate differentiation between differing yields. Regarded as an estimate of the mean yields of the fields actually sampled, the adjusted mean yield has a standard error of only $\sqrt{0.4137} = \pm 0.64$ bushels per acre.

* The closer approximation gives the value 0.733 .

7.15 Sampling with probabilities proportional to size of unit

In this case unbiased estimates of the sampling variances are those based on the mean square deviation of r . If s_r^2 is the estimated variance of r we have, for a random sample,

$$s_r^2 = S(r - \bar{r})^2 / (n - 1)$$

$$V(\bar{r}) = \frac{1}{n} s_r^2$$

If units selected more than once are included the number of times they are selected, no correction for finite population is required.

If the size of the population is known we therefore have

$$V(\bar{y}) = \bar{x}^2 s_r^2 / n$$

$$V(Y) = X^2 s_r^2 / n$$

If the size of the population is not known, the sampling variance of the estimate of total size X is derived from the formulæ for a qualitative variate (Section 7.4). We have, for a random sample, following the notation of Section 6.16,

$$V\left(\frac{n}{n_0}\right) = \frac{n}{n_0} \left(1 - \frac{n}{n_0}\right) \Big|_{n_0}$$

Hence, if A is known exactly,

$$V(X) = A^2 \frac{n}{n_0} \left(1 - \frac{n}{n_0}\right) \Big|_{n_0} = \frac{X^2}{n} \cdot \frac{n_0 - n}{n_0}$$

$$V(Y) = X^2 V(\bar{r}) + \bar{r}^2 V(X)$$

$$= \frac{X^2}{n} \left(s_r^2 + \frac{n_0 - n}{n_0} \bar{r}^2 \right)$$

If A is not known exactly its estimation will contribute some slight additional variance, the amount of which depends on the precise method of location of the points. This, however, will in general be sufficiently small to be neglected. Substituting $A = n_0/d$ for A , we have

$$V(X) = \frac{n(n_0 - n)}{d^2 n_0}$$

Example 7.15

Estimate the sampling errors of the estimates of Examples 6.16.a and 6.16.b, given that the standard deviation per field of the yield per acre is 3.5 cwt. per acre, and that the distribution of points can be taken as random.

The standard error of the mean yield per acre is

$$\text{S.E.}(\bar{r}) = 3.5 / \sqrt{529} = \pm 0.152 \text{ cwt.}$$

For the estimates of Example 6.16.a,

$$V(X) = 2560^2 \times \frac{529 \times 7788}{8317} \text{ acres}^2$$

$$S.E.(X) = \pm 57,000 \text{ acres}$$

$$V(Y) = \frac{1,354,000^2}{529} \left(3.5^2 + \frac{7788}{8317} 15.7^2 \right) = 84.24 \times 10^{10} \text{ cwt.}^2$$

$$S.E.(Y) = \pm 45,900 \text{ tons}$$

For the estimates of Example 6.16.b,

$$V(X) = 640^2 \times \frac{2202 \times 31,053}{33,255} = 842.2 \times 10^6 \text{ acres}^2$$

$$S.E.(X) = \pm 29,000 \text{ acres}$$

$$V(Y) = 1,409,000^2 \times 3.5^2/529 + 15.7^2 \times 842.2 \times 10^6 = 2536 \times 10^8 \text{ cwt.}^2$$

$$S.E.(Y) = \pm 25,200 \text{ tons}$$

Note that if the total area of crop were known accurately from other sources we should have

$$V(Y) = \frac{1,354,000^2}{529} 3.5^2 \text{ cwt.}^2$$

$$S.E.(Y) = \pm 10,300 \text{ tons}$$

If the acreage is not known a survey of this type will clearly be more efficient if sample harvesting is carried out at a proportion only of the sample points, the presence or absence of the crop being determined at the remaining points. This point is discussed in more detail in Section 8.17.

If the sample is such that it can be regarded as stratified by districts it might at first sight appear that the between-districts component should be eliminated from the variance of r . Unless the crop areas of the different districts are accurately known, however, this must not be done, since the proportion of sampling points falling in the crop in a district will not be accurately proportional to the area of the crop in that district.

If the sample points are confined to some localities only by a two-stage sampling process, with localities as first-stage sampling units, the above variances will represent the second-stage components of variance only. The full sampling errors must be determined from the first-stage units as explained in Section 7.17.

7.16 Sampling from within strata with probabilities proportional to size of unit

Since the number of units within each stratum is in general small, a pooled estimate s_r^2 of the within-strata variance of r based on the mean square

deviations of r within strata will be required. Consequently

$$s_r^2 = \frac{\sum S_i (r - \bar{r}_i)^2}{n - t}$$

We then have

$$V(\bar{r}_i) = s_r^2 (1 - f_i)/n_i$$

and thus

$$V(Y) = s_r^2 \sum \{X_i^2 (1 - f_i)/n_i\}$$

$$V(\bar{r}) = V(Y)/X^2$$

If the variation in the within-strata variances is large it may be necessary to introduce weights when forming the pooled estimate of error, in order to avoid bias.

The correction for finite sampling is not required if the units selected more than once are included the number of times they are selected. In the more usual case in which each unit is included once only, additional units being selected by the method of Section 3.10, the correction for finite sampling should be included. In this case, since probability of selection is not strictly proportional to size, the formulæ are approximate only.

Example 7.16

Estimate the sampling errors of the estimate of Example 6.17.

The analysis of variance of the values of the ratio for the individual "combined" parishes is given in Table 7.16. It follows the lines of Example 7.7.

TABLE 7.16—ANALYSIS OF VARIANCE OF THE VALUES OF THE RATIO

	Degrees of freedom	Sum of squares	Mean square
Between districts	6	·04952	·008253
Within districts	10	·02649	·002649
TOTAL	16	·07601	·004751

This gives a value of s_r^2 of 0·002649. We then have

$$\sum \{X_i^2 (1 - f_i)/n_i\} = 22,932^2 \times \frac{6}{7} + 43,591^2 \times \frac{13}{3} + \dots = 37 \cdot 188 \times 10^8$$

$$S.E.(Y) = \sqrt{(0 \cdot 002649 \times 37 \cdot 188 \times 10^8)} = \pm 3140$$

It will be noted that the variance of r within districts is less than the overall variance. Consequently stratification by districts appreciably reduces the sampling error.

7.17 Multi-stage sampling

If the sampling fraction at the first stage is small, the total sampling error of multi-stage sampling is obtained from the first-stage unit values, estimating each unit value from the results of the sampling at the second and following

stages, and using the method of estimation appropriate to the method of sampling at the first stage. The additional variability contributed by the second and following stages is automatically included in this estimate of error.

If, on the other hand, the sampling fraction f' at the first stage is not sufficiently small for the factor $(1 - f')$ to be neglected, the sampling error will be increased on account of the fact that the selected first-stage unit values are themselves subject to sampling error, instead of being known exactly, as in single-stage sampling. The increase in the sampling variance of whatever estimate is under consideration will be equal to f' times the variance in this estimate resulting from the sampling at the second and following stages. This variance can be calculated by regarding the selected first-stage units as strata which are sampled by the sampling at the second and following stages.

Thus, for example, in two-stage random sampling, with n' selected first-stage units, and n'' second-stage units selected from each first-stage unit, if s'^2 is the estimate of the sampling variance of the first-stage unit means, *i.e.* the means of the second-stage values for the separate selected first-stage units, s''^2 is the variance of the second-stage units about the first-stage unit means, and f' and f'' are the first- and second-stage sampling fractions, the sampling variance of the mean of the selected first-stage units will be $(1 - f'') s''^2 / n' n''$, since this mean is based on $n' n''$ second-stage units, and therefore the sampling variance of the mean of the population is

$$V(\bar{y}) = \frac{1 - f'}{n'} s'^2 + f' \frac{1 - f''}{n' n''} s''^2.$$

Example 7.17

Calculate the sampling errors of the estimates of the mean dressing of nitrogen per acre obtained in Example 6.19.

The y 's and x 's entering into the ratio method estimate at the first stage are the successive terms of the expressions for $S(g''y)$ and $S(g''x)$, already given for small farms, namely 1.36, 3.78, 3.30, . . . and 2, 6, 6, . . . respectively.

The calculation of the sampling error follows the lines indicated in Section 7.10, the ratio (cwt. nitrogen per acre) being assumed, for the reason given at the end of Section 7.9, to be the same for all strata.

The values of s_{qi}^2 are found to be

$$\begin{aligned} \text{Small farms:} & \quad 27.519/21 = 1.3104 \\ \text{Medium farms:} & \quad 583.81/35 = 16.680 \\ \text{Large farms:} & \quad 2779.1/8 = 347.39 \end{aligned}$$

We then have, neglecting the factors $(1 - f_i)$,

$$\begin{aligned} V(\bar{r}) &= (105^2 \times 22 \times 1.3104 + 59^2 \times 36 \times 16.680 + 30^2 \times 9 \times 347.39) / 58,229^2 \\ &= (\pm 0.0392)^2 \end{aligned}$$

The sampling fractions are here all small and the variance at the second stage therefore need not be considered. With the present material this variance could not in any case be estimated since only one field per farm was selected. In such cases, when the f_i are not small, it is still best to neglect them. The sampling error will then be slightly overestimated.

Farms without sugar beet have been excluded from the above calculation. Their inclusion, though substantially decreasing the values of s_{qi}^2 , would make little difference to the final estimate of error, since the values of n_i in the formula for $V(Y)$ would be correspondingly increased.

From inspection of the data, and from the nature of the material, we may expect that the variance of the mean dressing per acre r will be substantially constant, irrespective of the size of field to which it is applied. The alternative procedure of calculating the variance of r directly without any weighting may therefore be followed without serious risk of introducing any marked bias into the estimate of error.

TABLE 7.17—ANALYSIS OF VARIANCE OF DRESSINGS OF NITROGEN PER ACRE

	Degrees of freedom	Sum of squares	Mean square
Small farms	21	1.1304	0.0538
Medium farms	35	1.3500	0.0386
Large farms	8	0.9836	0.1230
TOTAL	64	3.4640	0.0541

The within-size-groups sums of squares and mean squares of r are shown in Table 7.17. There is no marked difference between the different size-groups, and in the following calculations we will therefore use the pooled estimate of the mean square, $s_r^2 = 0.0541$.

From the formulæ of Section 7.5 we have, for a single stratum,

$$V(\bar{r}_i) = s_{ri}^2 (1 - f_i) S_i (x^2) / \{S_i(x)\}^2$$

and

$$V(Y_i) = s_{ri}^2 (1 - f_i) S_i (x^2)$$

where the x 's are those entering into the first-stage sampling, i.e. $g''x$ in the full notation. The sums of squares of $g''x$ will be found to be 580, 15,245 and 39,993 for small, medium and large farms respectively. Consequently, summing over all strata as before, omitting the $(1 - f_i)$, and taking the variable sampling fraction into account, we have

$$\begin{aligned} V(\bar{r}) &= 0.0541 (105^2 \times 580 + 59^2 \times 15,245 + 30^2 \times 39,993) / 58,229^2 \\ &= (\pm 0.0391)^2 \end{aligned}$$

This is almost identical with the value previously obtained.

The comparison between the two methods of calculating the variance may be taken a stage further by estimating the values of s_{ri}^2 from those of s_{qi}^2 and

comparing them with those obtained directly. Equating the two expressions for $V(\bar{r}_i)$, we find $s_{ri}^2 = n_i s_{qi}^2 / S_i (g''x)^2$. Using the values of s_{qi}^2 already given we obtain for s_{ri}^2 the values 0.0497, 0.0394, 0.0782 respectively. These show no consistent divergence from the values of Table 7.17, and we may therefore conclude that the bias in the estimation of error by the second method is likely to be small. A more thorough investigation could be made by tabulating a number of comparisons of the above type from various batches of similar data.

The value of s_r^2 given by Table 7.17 is directly appropriate for the calculation of the variance of the estimate from the unweighted means (Table 6.19.c).

The sampling standard error of the mean dressing over all fields is $\sqrt{(0.0541/67)} = \pm 0.0284$. This standard error does not include any errors due to bias, but will be appropriate, or at least approximately so, to comparisons of such a nature that the major part of the bias is eliminated. If there were large differences between size-groups the question of whether the pooled within-size-groups variance or the overall variance is appropriate to the comparison in question would have to be considered—this, however, involves other problems, such as how far the differences observed are due to differences in size-group proportions (see examples 7.5 and 7.7.b).

It will be noted that the standard error of the properly weighted ratio estimate is considerably greater than the standard error of the straight mean. The ratio of the squares is $0.0392^2 / 0.0284^2 = 1.91$. Thus about double the number of farms, excluding those without sugar beet, are required to attain the same accuracy when unbiased estimates are required. This is inevitable in a survey of this kind where the sampling fractions cannot be adjusted so as to be proportional to the areas of the crop being sampled, a course which is impossible when a number of crops are covered in the same survey, even if the necessary information is available.

7.18 Systematic samples

No fully valid estimate of the sampling error of a systematic sample is possible, since the units are not located at random within defined strata. Approximate estimates can be made in various ways. The simplest, which will suffice for most census and survey work, is to divide the material arbitrarily into strata, and calculate the sampling error as if the units were selected at random from these strata.

In the case of a systematic sample from a list it will usually be sufficient to take account of the major groupings of the list, treating these as strata, and to ignore any minor and ill-defined groupings. An example of this has already been given (Example 7.7.a).

In the case of one-dimensional systematic sampling, *e.g.* equally spaced points on a line, or equally spaced lines covering an area, the strata may be taken to contain pairs of successive units, so that the error variance is estimated from the differences between the members of the pairs. Each difference

contributes one degree of freedom. If there are n' such differences d , the error variance per unit is therefore

$$s^2 = \frac{1}{2} S(d^2)/n'$$

Since the pairing is arbitrary, instead of taking alternate differences between successive units all differences may be taken. This is equivalent to taking two sets of overlapping strata. The accuracy of the estimate of s^2 is thereby somewhat increased, though it is not doubled, owing to lack of independence between the successive differences.

In the case of two-dimensional sampling on a square or rectangular pattern the strata should consist of sets of four units in a 2×2 pattern. By this means variability in both directions will be taken into account. There is no point in taking overlapping strata. Since each such stratum contributes 3 degrees of freedom the formula for the error variance per unit is

$$s^2 = [S(y^2) - \frac{1}{4} \Sigma \{S_i(y)\}^2]/3 n'$$

where n' is the number of strata.

In the case of line sampling a complication arises if all the lines are not of approximately equal length. If the total area covered by the sample is known, the most accurate estimate of the quantity under consideration will be obtained by the ratio method. In this case the calculation of the sampling error should strictly follow the method given in Section 7.9 for a stratified sample estimated by means of a constant ratio. With strata of two units the formula for Q becomes

$$Q = \frac{1}{2} S(dy^2) - 2 \bar{r} \cdot \frac{1}{2} S(dx dy) + \bar{r}^2 \frac{1}{2} S(dx^2)$$

This will eliminate the variability due to variation in length of line. If the total area is not known, so that the final estimate is obtained by multiplication of the total over all the lines by the appropriate raising factor, the difference method given above, and not the ratio method, must be used.

These methods of estimation of the sampling error are also applicable to line samples in which the lines are randomly located in pairs within blocks and thus form a proper stratified random sample. In this case the estimate of error will be fully valid.

In either systematic or stratified random line sampling, the variation in the length between neighbouring lines will not be large unless the boundaries of the area covered are very irregular. Consequently the approximate method based on the direct differences can be used in most cases without serious inaccuracy.

The above methods of estimation of error for systematic samples will give overestimates of the sampling error, provided there are no periodic features in the material, and provided in two-dimensional sampling that there are no marked strip effects running in straight lines across the material in such a manner that the whole of one line of sample points falls on the same strip. If a closer estimate is required, an alternative, but rather more complicated,

procedure is available. In one-dimensional sampling, instead of taking successive differences, differences of the type

$$d = \frac{1}{2}y_1 - y_2 + y_3 - y_4 + y_5 - y_6 + y_7 - y_8 + \frac{1}{2}y_9$$

can be taken. Such differences may be called *balanced differences*. Most of the systematic component of variation is thus eliminated. The number of terms included in each difference is to a certain extent arbitrary, but 9 is chosen as a convenient compromise. With extensive material there will be no need to take overlapping differences, the best procedure being to have overlap of the end terms only, so that the y_9 of the first difference is taken as the y_1 of the second. With this convention the sum of all the differences is equal to one-half the first and last included terms plus the sum of all the remaining odd terms minus the sum of all the even terms. The square of each difference contributes one degree of freedom, the divisor being given by the sum of the squares of the coefficients, *i.e.* 7.5. Consequently $s^2 = S(d^2)/7.5 n'$.

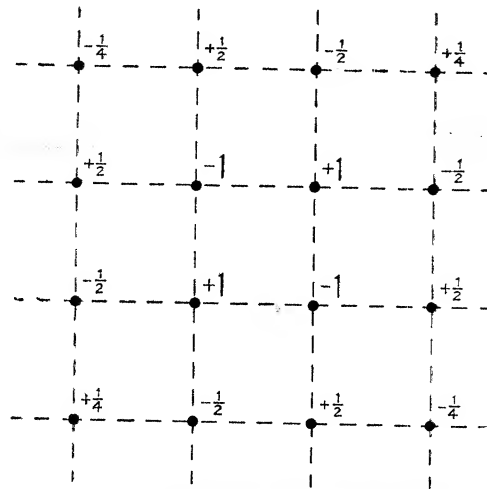


FIG. 7.18—COEFFICIENTS FOR CALCULATING THE ERROR OF A SYSTEMATIC TWO-DIMENSIONAL SAMPLE

A similar procedure can be followed in the case of two-dimensional systematic sampling, the most convenient type of difference being that given by the coefficients shown in Fig. 7.18. Here again, the margins of the square covering one difference may be taken as the margins of neighbouring squares. The divisor in this case will be $6\frac{1}{4}$.

The estimates provided by balanced differences will also in general be overestimates of the sampling error, but may be expected to be closer than those based on ordinary differences. If there is no wide discrepancy between the two types of estimate it may be concluded that the degree of overestimation is not likely to be great. More exact estimates can only be obtained by taking

supplementary observations at intermediate points allocated either at random or systematically. The one-dimensional case has been discussed in detail by Yates (1948, A).

The above methods of estimation can be applied both to quantitative and qualitative data, but in the case of qualitative data, based on either one- or two-dimensional point sampling, a rapid estimate of the sampling error can be made by using the formulæ for a random sample, as in Example 7.15. This will tend to give greater overestimation of the sampling error than the above methods, but if the parts of the line or area possessing the attribute are small and irregularly distributed, with no great variation in density in different parts of the line or area, the estimate will be sufficiently good for most practical purposes.

Example 7.18

In the 1942 Census of Woodlands the total area of woodland shown on the maps was determined for each county by estimating the area of land coloured green on the 1-inch O.S. maps. This was done by measuring the total length of the E-W kilometre grid lines which fell in green areas. The results for O.S. sheet No. 115 covering part of Kent are given in Table 7.18. Estimate the sampling error of this process.

TABLE 7.18—WOODLAND AREAS FROM LINE INTERCEPTS (cm.)

Grid line	Length of line, x	Length coloured green, y	Successive differences, d_y	Grid line	Length of line, x	Length coloured green, y	Successive differences, d_y
98	3.5	0.0	—	83	30.0	3.8	+ 1.4
97	4.2	0.9	+ 0.9	82	29.4	4.1	+ 0.3
96	9.2	0.0	— 0.9	81	29.1	4.9	+ 0.8
95	12.6	0.0	0.0	80	28.8	6.0	+ 1.1
94	15.5	0.3	+ 0.3	79	28.6	5.4	— 0.6
93	21.2	0.1	— 0.2	78	28.2	2.3	— 3.1
92	25.2	0.5	+ 0.4	77	27.2	2.9	+ 0.6
91	25.4	3.1	+ 2.6	76	26.3	2.1	— 0.8
90	31.2	2.8	— 0.3	75	25.4	6.3	+ 4.2
89	34.2	2.7	— 0.1	74	25.5	8.2	+ 1.9
88	34.1	2.8	+ 0.1	73	25.2	5.4	— 2.8
87	33.0	2.6	— 0.2	72	24.9	6.6	+ 1.2
86	31.4	2.3	— 0.3	71	24.6	6.6	0.0
85	31.0	3.5	+ 1.2	70	20.8	4.1	— 2.5
84	30.7	2.4	— 1.1				
					716.4	92.7	

The successive differences d_y of the lengths coloured green, y , are shown in the fourth and eighth columns. We find $S(d_y^2) = 63.21$ and consequently, since there are 29 lines,

$$s^2 = \frac{1}{2} 63.21/28 = 1.1288$$

$$\text{S.E. } \{S(y)\} = \sqrt{(29 s^2)} = \pm 5.72$$

A length corresponding to 1 km. represents an area of 1 sq. km. Consequently the raising factor to be applied to the total length measured in cm. to give the estimated area in acres is $63,360 \times 247.11/100,000 = 156.57$. The total area of woodland is therefore

$$156.57 \times S(y) = 156.57 \times 92.7 = 14,514 \text{ acres}$$

and the standard error of this area is $156.57 \times 5.72 = \pm 896$ acres, *i.e.* 6.2 per cent.

The same procedure can be followed for the other maps covering the county, and the square root of the sum of the squares of the resultant standard errors will give the standard error for the whole county, since the errors of the different maps are virtually independent. The results for Kent gave a percentage standard error of 3.4 per cent.

If the ratio method is used the corresponding successive differences d_x of the total lengths of the grid lines, and the sums of squares and products $S(d_x^2)$ and $S(d_x d_y)$ are required. The latter are found to be 159.23 and +2.62 respectively for the map in question. Using the ratio $\bar{r} = 0.12940$ derived from this map, we find

$$sq^2 = Q/28 = 1.1643$$

As expected, there is no appreciable difference in the error calculated by the two methods. The simpler method is consequently all that is really required, even when the total area of woodland is calculated from the total area of the county and the ratio of the length coloured green to the total length of the grid lines. If, however, the first grid line is much shorter than the rest, owing to its cutting the map boundary at a small angle, it should be omitted, or the length made up by taking the relevant part of the line on the neighbouring map. This trouble will only arise if the error is estimated separately for each map and the grid lines are not exactly parallel to the map boundary.

On the other hand, the calculation of the error from the total variance of y , *i.e.* without stratification, would give very misleading results.

7.19 Sampling on successive occasions

It will be sufficient if we record the variances of the estimates given in Sections 6.21 and 6.22.

(a) Two occasions only: sub-sample on the second occasion.

$$V(\bar{y}) = \{V(y) - \mu b^2 V(x)\} / \lambda n = (1 - \mu r^2) V(y) / \lambda n$$

$$V(\bar{y} - \bar{x}) = \{V(y) + (\lambda - 2\lambda b - \mu b^2) V(x)\} / \lambda n$$

If the population value on the second occasion is estimated by adding the estimate of change derived from the sub-sample only to the overall mean on the first occasion, *i.e.* by $\bar{y}' - \bar{x}' + \bar{x}$, the variance of this estimate will be

$$V(\bar{y}' - \bar{x}' + \bar{x}) = \{V(y) - \mu(2b - 1) V(x)\} / \lambda n$$

(b) Two occasions only : part of the sample replaced on the second occasion.

$$V(\bar{y}_w) = \frac{(1 - \mu r^2) V(y)}{n(1 - \mu^2 r^2)}$$

or, in the case of unequal numbers on the two occasions,

$$V(\bar{y}_w) = \frac{(1 - \mu r^2) V(y)}{n' + n''(1 - \mu r^2)}$$

With equal numbers on the two occasions, the variance of the estimate of change given by formula 6.21.b is approximately

$$V(\text{change}) = \frac{(1 - r) \{V(y) + V(x)\}}{n(1 - \mu r)}$$

The variance of the estimate given by the difference of the means of the units occurring on both occasions is approximately

$$V(\bar{y}' - \bar{x}') = (1 - r) \{V(y) + V(x)\} / \lambda n$$

and of that given by the difference of the overall means is approximately

$$V(\bar{y} - \bar{x}) = (1 - \lambda r) \{V(y) + V(x)\} / n$$

The exact expressions in the last two cases are given by replacing r by

$$2 \text{cov}(xy) / \{V(x) + V(y)\}$$

which is equal to r when $V(x) = V(y)$.

(c) Successive occasions : same fraction replaced on each occasion.

The limiting value of $V(\bar{y}_h)$, subject to the restrictions mentioned in Section 6.22, is as follows :—

$$V(\bar{y}_h) = \varphi V(y) / \mu n$$

The variance of the estimate of change given by $\bar{y}_h - \bar{y}_{h-1}$ is

$$V(\bar{y}_h - \bar{y}_{h-1}) = 2\varphi V(y) \{1 - r(1 - \varphi)\} / \mu n$$

Example 7.19.a

Estimate the sampling errors of the estimates of Example 6.21.

$V(x)$ and $V(y)$ may reasonably be taken as equal. The pooled estimate, based on all the observations on each occasion (22 degrees of freedom) is 0.08767.

We then have

$$V(\bar{y}_w) = \frac{0.08767 (1 - \frac{1}{3} \times 0.847^2)}{12 (1 - \frac{1}{3} \times 0.847^2)} = 0.0777^2$$

This may be compared with the variance of the overall mean $\bar{y}$, which is $0.08767/12$, i.e. 0.0855^2 . The ratio of these variances is 1.210. Thus the

gain in efficiency by the use of the information provided by the sampling on the first occasion is 21 per cent.

Similarly

$$V(\text{change}) = \frac{2 \times 0.08767 (1 - 0.847)}{12 (1 - \frac{1}{3} \times 0.847)} = 0.0558^2$$

The variance of the change estimated from the units sampled on both occasions is $2 \times 0.08767 (1 - 0.847)/8$, i.e. 0.0579^2 . The ratio of these variances is 1.077, and the gain is therefore 8 per cent. If the change is estimated from the difference of the overall means the variance is

$$V(\bar{y} - \bar{x}) = 2 \times 0.08767 (1 - \frac{2}{3} \times 0.847)/12 = 0.0798^2$$

The ratio of this variance to the first variance is 2.042, and the gain is therefore 104 per cent.

It will be noted that when $V(x)$ is taken as equal to $V(y)$ all the above gains depend solely on the values of λ and r .

Example 7.19.b

Estimate the sampling errors of the estimates of Example 6.22.

Owing to variation in the numbers on the different occasions the above formulæ will only give approximate estimates of the sampling errors. Excluding January, the average number of observations per occasion is 9.8 and the average value of λ is 0.664. Since $r = 0.811$, $1 - r^2 = 0.343$, and hence

$$\phi = \frac{-0.343 + \sqrt{[0.343 (1 - 0.657 \times 0.108)]}}{2 \times 0.664 \times 0.657} = 0.254$$

$V(y)$ was found to be 0.0871, and hence

$$V(\bar{y}_h) = \frac{0.254 \times 0.0871}{0.336 \times 9.8} = 0.0820^2$$

$$V(\bar{y}_h - \bar{y}_{h-1}) = \frac{2 \times 0.254 \times 0.0871 (1 - 0.811 \times 0.746)}{0.336 \times 9.8} = 0.0729^2$$

7.20 The error graph

In various instances in the preceding sections pooled estimates of the error variance have been used. Such pooled estimates are only legitimate if the error variance is reasonably constant over the parts of the population for which the pooling is carried out. In many types of material such constancy does not exist, and in such cases, when the number of degrees of freedom is too small for accurate determination of the error variances of the different parts of the population, other procedures must be followed if sampling errors are required separately for the different parts.

The simplest and most convenient device for practical use is the error graph. The estimates of the error variance are plotted against some other characteristic of the parts of the sample which is believed to govern the magnitude of the error variance, and a smooth curve is drawn to fit the points as closely as possible. This curve gives the variance law from which revised estimates of the variance can be obtained for any given value of the determining characteristic.

Fig. 7.20 shows a graph of this kind obtained in the course of a survey of wireworm infestation in grass fields. In each field 20 cores were taken and the wireworms counted in each core. There are thus 19 degrees of freedom for the determination of sampling error in each field. The estimates of the percentage variance so obtained were plotted against the estimated number of wireworms per acre in the various fields. The smooth curve so obtained was used to provide a table of errors that might be expected in similar sampling. Table 7.20 gives a small abstract of this table, and also of the inverse table, obtained by interpolation from the first table, giving the fiducial limits associated with any given observed number of wireworms. This procedure is approximate in a number of respects, but detailed discussion would be out of place here.

TABLE 7.20—DISTRIBUTION AND PROBABLE LIMITS OF ERROR OF SAMPLE ESTIMATES OF WIREWORM POPULATIONS OF GRASS FIELDS SAMPLED BY TWENTY 4 IN. CORES
(1,000 per acre)

True population	One-eighth of sample estimates		Estimated population	Population which, in one-eighth of cases, would give an estimate	
	less than	greater than		not less than that observed	not greater than that observed
200	105	295	200	128	325
400	260	540	400	284	567
600	428	772	600	451	804
800	597	1,003	800	624	1,040
1,000	766	1,234	1,000	797	1,277

There are, of course, various other methods of dealing with problems of this type. In biological work the original variates are often transformed to other variates, such as logarithms or square roots, which may in the material in question be expected to have a more constant variance, and thus permit pooling of the estimates of error. Such procedures introduce a number of

complications ; in particular the means of the transformed variates, when transformed back into the original variates, will be biased. They are not generally necessary or advisable in sample censuses and surveys.

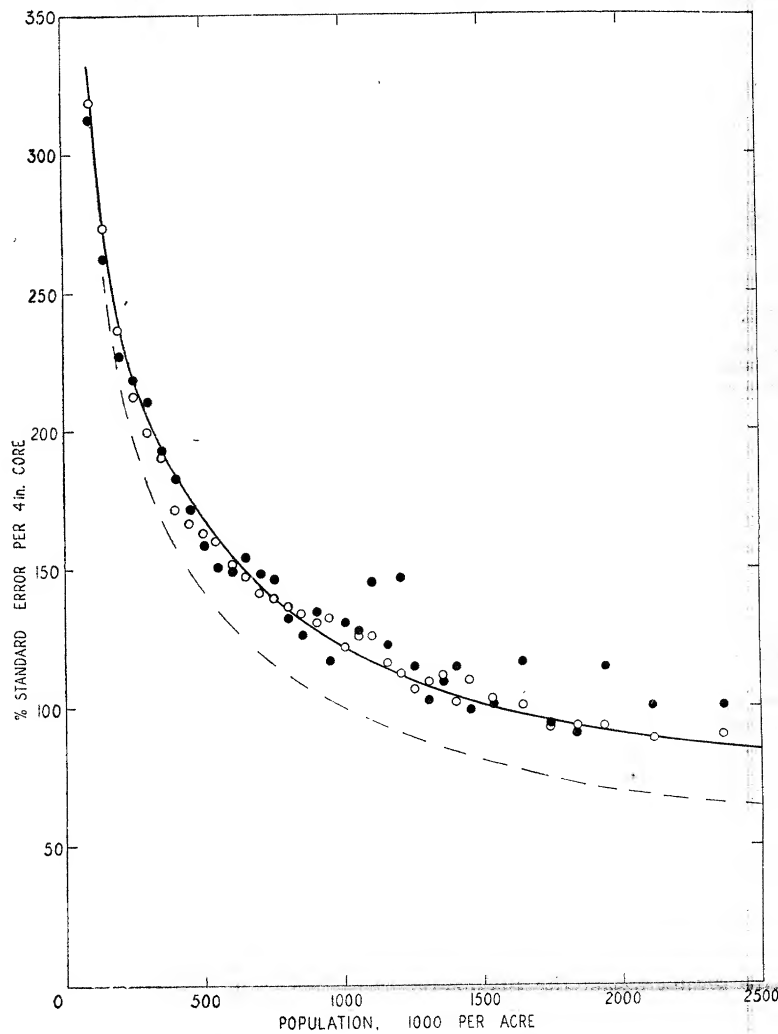


FIG. 7.20—STANDARD ERRORS PER UNIT CORE OF 4 IN. DIAM.
(WIREWORM SURVEY, 1940-1)

- means for 2272 fields grass in 1940
- means for 525 fields arable in 1940
- fitted to data from grass fields
- - - Poisson distribution

Reproduced from Yates and Finney (1942, J) with the permission of the editors of the *Annals of Applied Biology*.

7.21 Sub-sampling for the estimation of error

In an extensive survey the calculation of the sampling error from the whole of the data would be very laborious, and would provide estimates of error which are unnecessarily accurate. In order to cut down the work a sub-sample of the whole of the material may be taken, or estimates of error may be calculated for certain parts of the survey only, *e.g.* certain strata, with or without sub-sampling.

A convenient method of sub-sampling, which is applicable if there are a large number of separate strata of approximately equal size and a pooled estimate of the error variance per unit is required, is to select a random pair of units from each stratum, and to take the differences between the two members of each pair. In this case each difference d contributes one degree of freedom. If there are t differences the estimate of σ^2 is given by $S(d^2)/2t$.

If the strata are few in number and of unequal size this method is not applicable, since the number of differences would be inadequate and the different strata would not be represented in proportion to their size. In general it is important to see that the contributions to the error variance from the different parts of the population are substantially the same in the sub-sample as they would be if the whole of the data of the original sample were used. For this reason the sub-sample should in general be obtained by the use of a uniform sampling fraction over the whole of the original sample. A systematic method of selection will usually be satisfactory.

The taking of a sub-sample in this manner is somewhat troublesome, and also prevents accurate comparisons of the errors of parts of the survey which are in themselves small and therefore inadequately represented in the sub-sample. For these reasons the more convenient method of calculating the sampling error for certain parts of the population only is often employed. This procedure will lead to inaccuracies if the variability of the omitted portions is different from those that are included, but these inaccuracies can be reduced by selecting the parts to be included on a proper random basis. Thus in the 1942 Census of Woodlands the sampling error was calculated by selecting two counties at random from each of the seven regions, the data of the first 5 per cent. sample only being used. The surveyed quarter-sheets within each of these counties, which were selected on a systematic grid pattern, were treated as if they were a random sample from all the sheets of the county.

With grouped data the calculation of the sampling error from the whole of the data may well not present any appreciably greater labour than the use of a sub-sample, and in such cases the whole of the data will naturally be used.

7.22 Rounding-off and grouping errors

If a constant grouping interval over the whole of the range is adopted, the additional variance per unit introduced by the grouping is $\frac{1}{12}$ of the square

of the grouping interval. If the true variance per unit is σ^2 and the grouping interval is a , the total variance per unit of the grouped data is consequently $\sigma^2 + \frac{1}{12} a^2$, and the fractional loss of accuracy due to grouping is $a^2/12 \sigma^2$.

Rounding-off is a form of grouping. In terms of units of the last place that is retained, the variance per unit due to rounding-off is equal to $\frac{1}{12}$. If the variance is estimated from rounded-off or grouped data the additional variance due to grouping will be included in the estimate. If for any purpose an estimate of the true variance is required a deduction of $\frac{1}{12}$ the square of the grouping interval must be made. This is known as *Sheppard's correction*.

Grouping will also result in some loss of accuracy in the estimation of the variance. In a sample from a normal distribution the fractional loss of information due to this cause is $a^2/6 \sigma^2$.

The above formulæ are approximate, but provided the distribution is reasonably symmetrical they can be used in all cases ordinarily occurring in practice, in which a is not likely to be greater than, and is usually considerably less than, σ .

If the distribution is markedly skew, however, grouping may introduce a bias in the estimates which is not included in the above variance formulæ. An extreme case is provided by distributions in which there are a large number of small values and only few large values, such as the distribution of acres of crops and grass, or of wheat, in the farms of Table 6.6.a. In this type of distribution, if grouping is used, a smaller grouping interval must be employed for the small values than for the large values, as in Example 7.2.b. If only comparisons between similar estimates are required a coarser grouping may be adopted than if absolute values of the estimates are required, since in the former case any bias introduced will affect all estimates similarly.

Example 7.22

Determine the loss of accuracy in the estimation of the mean due to grouping in the data in Example 7.1.b.

The grouping interval is 300, and the standard deviation per unit (including grouping errors) is 526. Their ratio is 0.571, and the fractional loss of accuracy is therefore $\frac{1}{12} \times 0.571^2$ or 2.7 per cent., *i.e.* equivalent to 4 of the 162 families in the sample.

7.23 Determination of errors due to bias

As has been pointed out in Chapter 2, bias can arise either in the selection of the sample, or in the estimation process.

Although biased methods of estimation can in general be avoided, occasions arise, such as that discussed in Example 7.17, where biased estimates are considerably more accurate than the corresponding unbiased estimates. The biased estimates are also sometimes considerably simpler to calculate than the unbiased estimates. For these reasons it is sometimes advisable to use biased

estimates for comparative purposes. Such estimates can clearly be used with more confidence if the amount of bias that they introduce is not large. In general it is possible to make an estimate of the expected magnitude of a bias in estimation by a combination of mathematical analysis and detailed numerical analysis of the sampling results. Such methods, however, are complicated and vary with the type of sampling adopted. We shall therefore not describe them here.

An alternative and relatively simple method is to compare the biased estimates with the corresponding unbiased estimates of the same quantity. Any one comparison will of course be affected by random sampling errors in both the estimates, but if the material is sufficiently extensive to provide a number of comparisons, the mean difference will provide an estimate of the average bias whose accuracy can be judged by the variation of the individual differences.

Bias in the selection of the sample can in general only be assessed by comparison with another sample known to be free from bias. If, however, the distribution of some supplementary variate is known, bias in selection can sometimes be assessed, and if necessary eliminated, at least in part, by the use of regression. The calibration of eye estimates by means of regression, described in Section 6.15, provides an example of this procedure. As already pointed out, the procedure is subject to qualifications and cannot be relied on to compensate for all possible sources of bias. The only certain guarantee that bias is absent is the use of methods of selection and observation which are free from bias.

Example 7.23

Assess the evidence for bias in the estimate of the dressing of nitrogen per acre derived from the unweighted mean over all fields of Example 6.19.

The results already given in Tables 6.19.b and 6.19.c show that in each size-group the unweighted mean dressing is less than the weighted mean. The apparent bias in the overall unweighted mean dressing is -0.052 . A larger number of comparisons of the same type may be obtained by dividing the 22 farms of the small-size group into two groups of 11 farms each, and the 36 farms of the medium-size group into four groups of 9 farms each. Division of Table 6.19.a into blocks in this manner gives the comparisons shown in Table 7.23.

Six out of the seven differences are negative. The mean difference is -0.040 , and the standard error of this difference, estimated from the sum of the squares of the deviations of the individual differences from their mean, is ± 0.022 . The evidence for bias on this small amount of data is therefore not conclusive.

The procedure has here been adapted to the data given in Table 6.19.a. If the sampling were systematic the division of the groups should be made

by a random or systematic process such that each of the sub-groups constitutes a substantially random sample of the whole of the group. The procedure is also subject to the qualification that any bias arising from differential weighting of the different size-groups will be excluded by this method of estimation, since the differences are based on comparisons within size-groups.

TABLE 7.23—ESTIMATION OF BIAS

	Weighted mean	Unweighted mean	Difference
Small farms	.521	.484	— .037
	.378	.362	— .016
Medium farms	.584	.537	— .047
	.417	.446	+ .029
	.503	.470	— .033
	.378	.361	— .017
Large farms	.576	.416	— .160

An alternative method of subdivision which overcomes this limitation is to form sub-samples in which all the size-groups are represented in the correct proportions. If the data from a number of counties are available, no subdivision will be necessary, since the differences between size-group means and between the overall means for the different counties will provide all necessary comparisons.

7.24 Interpenetrating samples : comparison of observers

The error variance of the difference between two observers, estimated from interpenetrating samples, can be obtained by calculating the error variance appropriate to each observer, and adding these variances. This procedure, however, is subject to certain qualifications. In the first place the correction for finite sampling must not be applied. In the second place only those components of variance must be included which affect the comparisons between the observers. Thus, if a two-stage sampling process is adopted and each of the selected primary units is sampled by both observers, only the second-stage sampling error will enter into the comparison between observers.

If the data relevant to the comparison between two observers are at all extensive it is often possible to make a direct estimate of the error of this comparison by subdividing the material so that a number of independent differences are obtained, in the same manner as in the estimation of bias from different methods of estimation. Thus, with the above two-stage sampling process the difference between the observers might be obtained for each primary unit separately.

7.25 Estimation of the sampling error from duplicate samples

If a survey is carried out in two or more interpenetrating parts and the results are tabulated separately, an estimate of the sampling error can be obtained from the differences of the two samples. For such an estimate to be of any value there must be a number of independent differences, so that at least a moderate number of degrees of freedom are available. Even with extensive surveys the number of available differences is likely to be small, so that such estimates are usually rather rough. Nevertheless they are useful when the detailed results are not available.

If the two samples are distinguished by single and double dashes, and the estimate of the population total is given by the sum of the t parts 1, 2, etc., we have $Y' = Y_1' + Y_2' + \dots$, and $Y'' = Y_1'' + Y_2'' + \dots$. If the sizes of the two samples are in the ratio $\lambda : \mu$, where $\lambda + \mu = 1$, the estimate Y of the population total from the two samples is $\lambda Y' + \mu Y''$, with similar expressions for Y_1 , etc. An unbiased estimate of the error of Y is given by

$$V(Y) = \lambda\mu(1-f)\{(Y_1' - Y_1'')^2 + (Y_2' - Y_2'')^2 + \dots\}$$

where f is the sampling fraction for the whole survey. When the parts vary considerably in size this estimate is very inefficient, since excessive weight is given to the larger totals. If the approximate relation between the variances of the totals of the parts is known, a more efficient estimate can be obtained, though this will be biased if the assumed law of variance is incorrect. If the variances are proportional to Y_1, Y_2 , etc., the efficient estimate is

$$V(Y) = \frac{Y\lambda\mu(1-f)}{t} \{(Y_1' - Y_1'')^2/Y_1 + (Y_2' - Y_2'')^2/Y_2 + \dots\}$$

This law of variance is likely to be approximately true for area surveys if the density per unit area of the quantity surveyed does not vary very greatly from part to part.

If the variances are proportional to some other quantity, such as the number of units in each part, these numbers must be substituted for Y_1, Y_2 , etc. and their sum for Y in the above formula.

Example 7.25

In the 1942 Census of Woodlands the total volumes of timber for the seven regions of the survey obtained from the first and second 5 per cent. samples, excluding areas surveyed in 1938-9, and with allowance for felling in the interval between the two samples, are shown in Table 7.25. Estimate the sampling error to which the combined estimate of the total volume of timber for the country is subject.

TABLE 7.25—VOLUMES OF TIMBER IN THE DIFFERENT REGIONS ESTIMATED FROM THE FIRST AND SECOND 5 PER CENT. SAMPLES OF THE 1942 CENSUS OF WOODLANDS

Region	Volume, m. cu. ft.			$A - B$	$\frac{(A - B)^2}{\text{Mean}}$
	Sample A	Sample B	Mean		
A	253	201	227	+ 52	11.9
B	74	84	79	- 10	1.3
C	100	107	104	- 7	0.5
D	148	164	156	- 16	1.6
E	209	227	218	- 18	1.5
F	94	78	86	+ 16	3.0
G	112	119	116	- 7	0.4
	—	—	—	—	—
	990	980	986	+ 10	20.2

The computations are shown in the last three columns. The sum of squares of the differences $A - B$ is 3738, and therefore from the first formula the standard error of the total is

$$\sqrt{\left(\frac{1}{2} \times \frac{1}{2} \times \frac{9}{10} \times 3738\right)} = \pm 29.0 \text{ m. cu. ft.}$$

The sum of the last column is 20.2, and therefore by the second formula the standard error is

$$\sqrt{(986 \times \frac{1}{2} \times \frac{1}{2} \times \frac{9}{10} \times \frac{1}{7} \times 20.2)} = \pm 25.3 \text{ m. cu. ft.}$$

It will be noted that an estimate of the standard error of any regional total can be obtained directly from the second formula by substituting this total for the grand total 986.

The above estimates are very rough, since they are based on only 7 degrees of freedom. The estimate obtained by the method described in Section 7.18 is ± 19.7 m. cu. ft. With a more extensive set of comparisons between the two samples a more accurate estimate could be obtained.

7.26 Presentation of sampling errors in extensive surveys

In the preceding sections the methods of estimating the sampling error of single estimates, *e.g.* of the population mean, have been described. In extensive surveys the results will usually be broken down in various ways,

so that the final tables will contain a large number of estimates relevant to different parts of the population. The errors of these different estimates could be calculated separately by the methods already given, using pooled estimates of the error variance per unit where appropriate, but such a procedure would be laborious, and even if it were carried out, the presentation of the separate standard errors in the tables of results would make these tables somewhat confused.

From every point of view, therefore, it is desirable to have some condensed method of presenting the errors of the component parts of elaborate tables. This can be effected by making use of some error law, either theoretical or empirical, which will enable the standard error of any particular component to be rapidly obtained when required from other information available in the table. The exact form of the law will depend on the type of material and the nature of the information tabulated.

The simplest case is that in which the tables contain means of a quantitative variate derived from a survey with constant sampling fraction, and the error variance per sampling unit is constant. The standard error of any mean then depends only on the value of the sampling fraction, the error variance, and the number of units on which the mean is based. These numbers are likely in any case to be of interest in themselves and to be presented either directly or in raised form as estimates of the numbers of units in the different parts of the population. Alternatively, if the actual numbers of units in the different parts are known these may be presented.

Whatever the exact form of presentation an auxiliary table can easily be prepared by the use of formula 7.2, or its analogue for stratified sampling, giving the standard errors corresponding to different numbers in the population or sample. If a table is felt to be unduly elaborate the formula on which it is based may be presented. If the standard errors are likely to be used mainly for testing the differences between different means the correction for finite sampling can be omitted, with of course a note to this effect. It may be noted that with a table of this type the standard error of the difference of two means based on numbers n_1 and n_2 can be obtained by entering the table with the number n' , given by $1/n' = 1/n_1 + 1/n_2$.

Another simple case is that in which the summary table relates to qualitative data based on the proportion of units possessing a given attribute. If the sample is random the standard error of any entry depends only on the sampling fraction, the proportion of units possessing the given attribute in the part of the population under consideration, and the number of units. If the variation in the proportion of units is not large in the different parts of the population, a table or formula based on the proportion in the whole population may be sufficiently accurate. If the proportion of units possessing the attribute is small in all parts, q can be taken equal to unity.

An example of the use of an empirical law is provided by the case in which the error variances are approximately proportional to the magnitude of the totals of the different parts, which, as pointed out in the last section, is likely

to hold for certain types of area survey. In this case a formula or table relating the errors to the entries of a table of totals can be presented.

There is not space here to discuss more complicated cases, which must be dealt with on their merits as they arise. With the more elaborate types of sampling the possibilities for presenting the standard errors in the form of auxiliary tables are more limited, but even in such cases it is often possible to summarize the standard errors in the form of a few relatively simple formulæ, suitable for rapid calculation on a slide rule.

EFFICIENCY

8.1 General remarks

The methods described in Chapter 7 enable the sampling error associated with a sample of a given type and size to be calculated from the data furnished by the sample itself. When planning a sample census or survey, we have to solve the more general problem of calculating the sampling errors of samples of various types and sizes from the data furnished by a sample of a particular type and size. We can then determine which method of sampling is likely to be most efficient and the size of sample necessary to give the required accuracy.

The determination of the sample size in the case of a random sample from a large population has already been discussed in Section 4.31. It was there shown that, for qualitative characters which are attributes of the sampling units, the number of units required could be determined without any prior knowledge of the material other than the approximate proportion of units possessing the given attribute in the population; and that for quantitative characters knowledge of the standard deviation of the character in question per sampling unit was all that was required.

The formulæ of Section 4.31 apply when the population is large relative to the size of sample required. If the population is not large a correction must be made to allow for finite sampling. This is most simply done by calculating the number of units n_0 that would be required if the population were large, and the corresponding sampling fraction $f_0 = n_0/N$. The required sampling fraction is then given by

$$f = \frac{f_0}{1 + f_0} \quad (8.1)$$

In this calculation f_0 may be greater than unity.

The method followed in Section 4.31, *i.e.* that of taking the appropriate formula for the standard error of a sample of size n and rewriting this formula to give an equation for n , is a general one and can be applied to the more complicated types of sample, using the appropriate formulæ for the standard errors given in Chapter 7. It is apparent, however, that these formulæ can only be used if the relevant variances per sampling unit are known or can be estimated. In certain cases, also, the formulæ cannot conveniently be rearranged so as to give n directly. This, however, is a minor point, since the required solution can always be quickly found by trial once estimates of the relevant variances are available.

In the following sections we will discuss the problems that arise in the estimation of the variances relevant to different types of sample when the basic data consist of a sample of a different type. In certain cases data relating

to all the units of a population will be available. This situation does not differ in any essential particulars from that in which data are derived from a random sample of the population.

We will here define the sense in which we shall use various terms in the subsequent discussion.

The *relative accuracy* of two samples which differ in respect of method of sampling or size of sample, or both, may be defined as the reciprocal of the ratio of the sampling variances of the estimates provided by them.

The *relative precision* of two different methods of sampling based on the same type of sampling unit may be defined as the reciprocal of the ratio of the sampling variances of the estimates given by the two methods when the same number of units are taken.

The *relative efficiency* of two different methods of sampling based on the same type of sampling unit may be defined as the reciprocal of the ratio of the numbers of units required to attain a given accuracy with the two methods.

In the case of a random sample from a large population, or a stratified sample with fixed strata from such a population, the relative efficiency is equal to the relative precision. But if the size of the strata depends on the number of units in the sample, or if the population is not large relative to the size of the sample, there is a difference between the two concepts.

The term *efficiency* is already in current use in the theory of estimation. It is there used in an absolute sense. An estimate is *efficient* (*i.e.* has an efficiency of 100 per cent.) if in large samples it is one of the class of most accurate estimates, *i.e.* estimates with minimum variance. An estimate has an efficiency of x per cent. if it has $100/x$ times this minimum variance. This use of the term is analogous to *precision* in our terminology. The reason why no distinction has to be made between precision and efficiency in the theory of estimation is that only large populations are normally under consideration, in which case the two concepts are synonymous. Since no confusion is likely to arise, we shall continue to use the term efficiency when discussing the relative accuracy of different estimates derived from the same sample.

The concepts of relative precision and relative efficiency may be extended to cover methods of sampling based on different types of sampling unit, by replacing numbers of units by the amount of material included in the sample. They may be further extended to cover the relative accuracy for a given cost and the relative cost for a given accuracy.

It may be noted here that the relative precision and relative efficiency of different types of sampling should as far as possible be judged from estimates of the sampling variances derived from the same set of data. Comparisons based on estimates derived from independent samples of different types are subject to errors of estimation which are considerably larger, and comparisons based on samples from different aggregates of similar material are even more subject to uncertainty. No very general conclusions should, however, be drawn from a single comparison based on a small amount of data, even when a single set of data is used. The relative precision of stratified and random

samples, for instance, will depend on the differences between strata, and these differences may vary considerably even in apparently similar material.

8.2 Qualitative data

If the variates under consideration are attributes of the sampling units, the effect of stratification, with either uniform or variable sampling fraction, can be determined from a knowledge of the proportions of units possessing the given attribute in the different strata. In other cases qualitative variates must be treated similarly to quantitative variates, as in the estimation of sampling errors.

Formulae for the required size of a stratified random sample with uniform sampling fraction, analogous to those for a random sample given in Section 4.31, can be written down without difficulty. A somewhat simpler approach, however, is to estimate the percentage standard error of a stratified sample of any convenient size (*e.g.* the size of the sample of which the data are available) on the assumption that the population is large. The size of sample required to give any predetermined percentage standard error is then given, if the population is large, by the formula

$$\frac{\text{Size of sample required}}{\text{Size of actual sample}} = \frac{(\text{Actual percentage standard error})^2}{(\text{Required percentage standard error})^2}$$

Allowance for the effect of finite population size can then be made by formula 8.1.

In the case of a stratified random sample with variable sampling fraction the same procedure can be followed, with the exception that allowance for the effect of finite population size cannot be made in the above manner. If, therefore, any of the correction factors $(1 - f_i)$ are sufficiently large to be of importance, the approximate size of sample required may first be calculated as above and the final size found by trial. Variable sampling fractions, however, are not likely to be much used for qualitative data.

Example 8.2.a

If a large population of individuals is divided into five strata containing equal numbers of people, determine the relative sizes of a stratified and a fully random sample of the same accuracy when the percentages of individuals giving a positive answer to a given question in the different strata are (a) 70, 60, 50, 40 and 30 per cent, (b) 10, $7\frac{1}{2}$, 5, $2\frac{1}{2}$ and 0 per cent.

A sample of 500 people will have 100 in each stratum. The variance of the number in the sample giving positive answers will be, in case (a), for a stratified sample,

$$100 \times .7 \times .3 + 100 \times .6 \times .4 + 100 \times .5 \times .5 + 100 \times .4 \times .6 + 100 \times .3 \times .7 = 115$$

and for a random sample,

$$500 \times .5 \times .5 = 125$$

The ratio of the required sizes is therefore $125/115 = 1.087$, *i.e.* the random sample will have to be 8.7 per cent. larger. In case (b) a similar calculation shows the random sample will have to be 2.7 per cent. larger.

Example 8.2.b

Determine from the data of Examples 6.5 and 7.6.a the numbers of farms required to give a sampling standard error of 5 per cent. in the estimate of the number of farms growing wheat (a) when the sample is random, (b) when it is stratified by size-groups.

(a) We have $p = 54/125 = 0.432$. Consequently, from Sections 4.31 and 8.1

$$n_0 = \frac{10,000 \times 0.568}{0.432 \times 5^2} = 526$$

$$f_0 = 526/2496 = 0.210$$

$$f = \frac{0.210}{1 + 0.210} = 0.174$$

$$n = 433$$

(b) We have already found that for the stratified sample $U = 1080$ and $S.E.(U) = \pm 71.64$. If the population had been large, therefore, we should have had $S.E.(U) = 71.64/\sqrt{(1 - 1/20)} = 73.5$. Consequently in this case $S.E. \% (U) = 6.80$. Hence

$$\frac{n_0}{125} = \frac{6.80^2}{5^2}$$

$$n_0 = 231$$

$$f_0 = 231/2496 = 0.0927$$

$$f = 0.0927/(1 + 0.0927) = 0.0848$$

$$n = 212$$

The standard error of the total estimated from a random sample of 125 farms is $20\sqrt{(125 \times 0.432 \times 0.568 \times 19/20)} = 108.0$. Consequently the relative precision of the stratified and random samples of 125 units (or indeed of any number of units) is given by $108.0^2/71.64^2 = 2.27$. The relative efficiency, when a 5 per cent. standard error is required, however, is $433/212 = 2.05$. The relative efficiency is slightly less than the relative precision because we are sampling from a finite population.

8.3 Random sample and stratified sample with uniform sampling fraction

The general principle to be followed is to construct an analysis of variance which corresponds as closely as possible to that appropriate to the required

type of sample. The procedure varies somewhat according to the type of data available.

(a) From the data of a stratified sample with uniform sampling fraction :

An analysis of variance within and between strata in the form of Table 7.7.a must be made. The within-strata mean square s_1^2 gives an estimate of the error variance per unit in a stratified sample, and the mean square s^2 from the total line gives a similar estimate for a random sample. If separate estimates of the error variance per unit have been made for the different strata, as in Example 7.6.a, a pooled within-strata sum of squares may be calculated by multiplying the within-strata error variances by the degrees of freedom $n_i - 1$ for each stratum, and summing the products, or by summing the sums of squares directly.

The formula of Section 4.31 can then be used to determine the size of sample, using s_1^2 in place of s^2 for a stratified sample, and correcting for finite population in the same manner as in Section 8.1. Since for a stratified sample $V(\bar{y}) = s_1^2(1-f)/n$, and for a random sample $V(\bar{y}) = s^2(1-f)/n$, the relative precision of stratified and random sampling will be given by the ratio of s^2/s_1^2 . The relative efficiency will be somewhat less than the relative precision when the corrections for finite sampling are appreciable.

This procedure is approximate in two respects. In the first place, if the variances within the different strata are unequal they do not enter into the mean square B with quite the correct weights, as already explained in Section 7.7. In the second place, a stratified sample has a slightly greater overall variance per unit than a random sample from the same population, and consequently C is not the best estimate of the variance per unit of a random sample. Neither of these approximations gives rise to errors of any importance in the comparison of a random and a stratified sample, but it may be noted that the bias in C can be almost completely eliminated by calculating s^2 from the formula

$$s^2 = \{(n-1)C + B\}/n \quad (8.3)$$

An extension of this formula is of use in the case of multiple stratification (Section 8.4). Method (c) below takes account of both sources of disturbance.

(b) From the data of a random sample :

An analysis of variance within and between strata can be made in the same manner as with a stratified sample with uniform sampling fraction, and s_1^2 and s^2 can be estimated as in a stratified sample.

For this procedure it is only necessary that the units of the sample be classified by strata. The numbers of units of the whole population falling in the different strata do not require to be known.

If these numbers are known, Method (c) below can be followed. This will give slightly more accurate results at the cost of a little additional computation, since allowance is made for the fact that the numbers in the

different strata in the sample will not be exactly proportional to the numbers in the population, owing to the fluctuations of random sampling.

(c) From the data of a stratified sample with a variable sampling fraction (or any arbitrary values of the sampling fractions):

Estimates of the average within-strata mean square s_1^2 and of the overall mean square must be calculated from the proportions $h_i = N_i/N$ of the units of the population in the different strata. The formulæ are

$$s_1^2 = \sum h_i s_i^2$$

$$s^2 = s_1^2 + \sum h_i \bar{y}_i^2 - \bar{y}^2 - \sum h_i (1 - h_i) s_i^2 / n_i$$

where $\bar{y}$ is the estimate of the population mean derived from the sample, and is consequently equal to $\sum h_i \bar{y}_i$. The relation of these formulæ to the analysis of variance of a stratified sample with uniform sampling fraction will be apparent. The terms involving y in s^2 correspond to the between-strata component of variance, the last term of s^2 being the correction required because the $\bar{y}_i$ are themselves subject to sampling error. This correction will be trivial except when the between-strata component of variance is small and there are a large number of strata with few units from each stratum. If the h_i are put equal to n_i/n (uniform sampling fraction), s^2 will be the same, to order $1/n$, as that given by the mean square C of Method (a), with the exception that in Method (a) $\sum h_i \bar{y}_i^2 - \bar{y}^2$ is multiplied by a factor $n/(n-1)$.

It will be noted that the data need not be derived from a sample in which the sampling fractions are chosen with the object of obtaining the most accurate possible estimates: any set of data in which the sampling is random within strata, and from which the proportions of the units in the different strata, the strata means and the within-strata variances can be determined with sufficient accuracy, will be adequate.

Example 8.3.a

Determine the error variances per unit and the relative precision of a stratified random sample with uniform sampling fraction and a fully random sample from the data on wheat acreages of the stratified random sample of Hertfordshire farms (Examples 6.5 and 7.6.a).

The analysis of variance is given in Table 8.3.a. The within-strata sum of squares is obtained directly from Table 7.6.a by summing the column $S_i (y - \bar{y}_i)^2$. The between-strata sum of squares is obtained by summing the products of the columns of Table 6.5.b giving the totals and means, and deducting the product of the general total and the general mean. These means should be taken to two and three decimal places respectively. We thus have $s_1^2 = 349.2$ and $s^2 = 797.1$. The estimate of the relative precision is therefore 2.28.

TABLE 8.3.a—ANALYSIS OF VARIANCE OF THE STRATIFIED RANDOM SAMPLE OF
HERTFORDSHIRE FARMS

	Degrees of freedom	Sum of squares	Mean square
Between size-groups .	5	57,278	
Within size-groups .	119	41,558	349.2
Whole sample . .	124	98,836	797.1

Example 8.3.b

Make similar estimates to those of Example 8.3.a, using the data of the random sample (Examples 6.6, 7.2.b and 7.6.b).

The analysis of variance is given in Table 8.3.b. If the N_i are not known, we have $s_1^2 = 488.5$ and $s^2 = 1329.9$. The estimate of the relative precision is therefore 2.72.

TABLE 8.3.b—ANALYSIS OF VARIANCE OF THE RANDOM SAMPLE OF HERTFORDSHIRE
FARMS

	Degrees of freedom	Sum of squares	Mean square
Between size-groups .	5	106,775	
Within size-groups .	119	58,129	488.5
Whole sample . .	124	164,904	1,329.9

If the N_i are known the calculations follow the same lines as those of Example 8.3.c below, and are left to the reader. In this case we find $s_1^2 = 436.5$ and $s^2 = 1189.2$, the estimate of the relative precision being again 2.72.

Example 8.3.c

Make similar estimates to those of Example 8.3.a, using the data of the sample with variable sampling fraction (Examples 6.7 and 7.7.a).

Table 8.3.c shows the calculations. The h_i are calculated from the numbers in the population. These are given in Table 6.6.b, except for the last two size-groups, which have the values 215 and 51 respectively. It will be noted that we are here considering a sample stratified for districts as well as size-groups.

TABLE 8.3.c—CALCULATION OF THE AVERAGE WITHIN-STRATA AND OVERALL MEAN SQUARES FROM THE STRATIFIED SAMPLE WITH VARIABLE SAMPLING FRACTION OF HERTFORDSHIRE FARMS

Size-group	h_i	n_i	$\bar{y}_i$	$\bar{y}_i^2$	s_i^2	$h_i(1-h_i)s_i^2/n_i$
1- 5	.174	0	(0)	(0)	(0)	(0)
6- 20	.208	3	0	0	0	0
21- 50	.143	6	4.5	20	53.5	1.1
51-150	.208	26	8.2	67	159.2	1.0
151-300	.160	40	29.1	847	564.2	1.9
301-500	.086	43	76.6	5,868	1,703	3.1
501-	.020	17	172.1	29,618	2,614	3.0
	.999	135	17.03	1,249.3	329.8	10.1
301-500	.811	43	76.6	5,868	1,703	6.1
501-	.189	17	172.1	29,618	2,614	23.5
	1.000	60	94.65	10,357	1,875	29.6
301-	.106	60	94.6	8,959	3,243	5.1
	.999	135	17.03	1,102.0	474.8	9.1

The sums of the products of h_i with $\bar{y}_i$, $\bar{y}_i^2$ and s_i^2 are shown at the foot of their respective columns. We therefore have, since $17.03^2 = 290.0$,

$$s_1^2 = 329.8$$

$$s^2 = 329.8 + 1249.3 - 290.0 - 10.1 = 1279.0.$$

Hence the relative precision is $1279.0/329.8 = 3.88$. It will be seen that the corrections in the last column are here trivial, and could well be omitted.

Size-groups 301-500 and 501- can be combined in the manner shown in the second part of Table 8.3.c. We have, for these two size-groups combined,

$$s^2 = 1875 + 10,357 - 8959 - 30 = 3243$$

We can now insert a fresh line in Table 8.3.c to replace the lines for the last two size-groups in the first part of the table. The previous computation is then repeated, giving

$$s_1^2 = 474.8$$

$$s^2 = 474.8 + 1102.0 - 290.0 - 9.1 = 1277.7$$

Hence the relative precision is 2.69.

The amalgamation of the two size-groups containing the largest farms has resulted in a considerable loss of precision, the relative precision being $329.8/474.8 = 0.69$.

8.4 Multiple stratification

The gain in precision due to sub-stratification of a sample which is already stratified into main strata can be estimated by methods similar to that of Section 8.3. An example has already been given in Example 8.3.c, where the gain in precision resulting from the subdivision of the size-group 301– into two groups, 301–500 and 501– was determined.

If the data are derived from a sample with uniform sampling fraction which is itself sub-stratified, the comparisons can be made directly between the relevant mean squares in the analysis of variance, as in Method (a). The structure of the analysis of variance in this case is

$$\text{Whole sample } (s^2) \begin{cases} \text{Between main strata} \\ \text{Within main strata } (s_1^2) \begin{cases} \text{Between sub-strata} \\ \text{Within sub-strata } (s_2^2) \end{cases} \end{cases}$$

The ratio of the mean squares s_1^2 and s_2^2 within main strata and within sub-strata will give the required relative precision.

A similar analysis can be constructed from data derived from other types of sample with uniform sampling fraction (Method (b)).

One case of practical importance is that in which both the main and sub-strata are arbitrary subdivisions of an area, all the main and all the sub-strata being of equal size. If there are t' main strata, and t'' sub-strata per main stratum, with k selected sampling units per sub-stratum, the analysis of variance will be of the form shown in Table 8.4.

TABLE 8.4—STRUCTURE OF THE ANALYSIS OF VARIANCE IN A DOUBLE STRATIFICATION

	Degrees of freedom	Mean square
Between main strata	$t' - 1$	A
Within main strata $\begin{cases} \text{Between sub-strata} \\ \text{Within sub-strata} \end{cases}$	$\begin{cases} t' (t'' - 1) \\ t' t'' (k - 1) \end{cases}$	$\begin{cases} D \\ E = s_2^2 \end{cases}$
Total	$t' t'' k - 1$	$B = s_1^2$
Total for sample	$t' t'' k - 1$	$C = s^2$

If k is small the bias in the estimate s_1^2 provided by the within-strata mean square may be appreciable. This bias can be almost completely eliminated by using the formula

$$s_1^2 = \{(t'' k - 1) B + E\} / t' k$$

which is derived directly from formula 8.3.

8.5 Stratified sample with variable sampling fraction

In the notation of Section 8.3, we have

$$N V(\bar{y}) = \sum s_i^2 h_i (1 - f_i) / f_i$$

The n_i or f_i required for a given accuracy can only be determined uniquely from this equation if the relations between the different f_i have been decided.

It has already been pointed out (Section 3.5) that for maximum accuracy the f_i should be proportional to σ_i , but that in many types of material stratified by size-groups the f_i may be taken proportional to the mean sizes of the size-groups. If we put $f_i = c \lambda_i$, where the λ_i are in the required proportions, the above equation can be written

$$\frac{1}{c} \sum (s_i^2 h_i / \lambda_i) = N V(\bar{y}) + \sum s_i^2 h_i$$

The value of c for any required accuracy can then be calculated. If, however, a value of c is obtained which makes some of the f_i greater than 1 the calculation must be repeated, omitting the terms for these strata from both sides of the equation.

Alternatively the direct expression for $V(\bar{y})$ can be used and the value of c found by trial. This has the advantage that the effect of adjustments of the final sampling fractions to simple fractions is immediately apparent.

The relative precision of stratified samples with variable and with uniform sampling fractions can be obtained by calculating $V(\bar{y})$ for both samples. It should be noted that if the f_i have been taken proportional to the s_i a slight over-estimate of the relative precision will be obtained, owing to errors in the s_i . This point has been discussed by Sukhatme (1935, A), but is not of great importance in practice.

It will be seen that for these calculations we only require sufficiently accurate estimates of the variances within strata and the proportions of units of the population in the different strata. The procedure is therefore the same whether or not the sample from which the data are obtained is stratified. All that is required is that all strata should be adequately represented.

Example 8.5.a

From the data of Table 8.3.c determine the size of sample required to give a standard error of ± 1500 acres in the estimate of wheat acreage, when sampling fractions proportional to those of Table 3.7.a are used.

The λ_i can be taken equal to the sampling fractions of Table 3.7.a. Tabulating $s_i^2 h_i$ and $s_i^2 h_i / \lambda_i$, we find

$$\sum (s_i^2 h_i / \lambda_i) = 2913 \qquad \sum s_i^2 h_i = 329.77$$

Also $N V(\bar{y}) = V(Y)/N = 1500^2/2496 = 901.44$, and hence

$$c = 2913 / \{901.44 + 329.77\} = 2.37$$

The total number required in the sample is therefore $135 \times 2.37 = 320$, the number in the largest size-group, for example, being $51 \times 2.37/3 = 40$. No sampling fraction is greater than 1, and therefore no further computation is required.

In practice the new sampling fractions may well be rounded off, taking, for example, all of the largest size-group, $\frac{1}{2}$ of the next, etc.

The direct approach illustrates the way in which results of this kind can readily be obtained by interpolation. A first approximation to the number required is given by $135 \times 2550^2/1500^2 = 390$. The standard error corresponding to this sample number, obtained by the ordinary methods, is ± 1302 . The squares of the reciprocals of this and of the original standard errors can be plotted against the respective numbers in the samples and a smooth curve drawn through these two points and the origin. This curve gives the general relation between sample size and accuracy, and will be found to give a sample number corresponding to a standard error of ± 1500 of approximately 320.

Example 8.5.b

Determine the relative precision of the sample of Table 3.7.a and the sample with sampling fractions proportional to s_i containing the same number of farms.

The standard error, using these sampling fractions, can be calculated in the ordinary manner, and is found to be ± 2420 . The relative precision is therefore $2420^2/2550^2 = 0.90$. There is consequently an apparent loss of precision of approximately 10 per cent., but the real loss is likely to be less than this, owing to errors in the estimates of the standard errors.

This apparent loss refers to a single variate, acreage of wheat. If, for instance, the acreage of some other crop were taken, the σ_i would be different and the sampling fractions required to give minimum variance would therefore also be different. Consequently, if several variates have to be determined, a compromise will in any case be required.

8.6 Supplementary information

The determination of the number of units required in a sample when supplementary information is available presents no essentially new problems. It has been shown in Chapter 7 that apart from the substitution of s_q^2 or s_l^2 for s^2 the formulæ for the variances of estimates based on supplementary information differ little from those for estimates from similar samples without supplementary information. Consequently it will usually be sufficient to estimate the appropriate variance by the methods given in Chapter 7, using this variance instead of the ordinary variance per unit to determine the size of sample. The factor $\bar{x}^2/\bar{x}^2$ in the variance of the ratio estimate differs from unity only because of sampling fluctuations in $\bar{x}$, and can be omitted.

When the ratio method is to be used and $V_x(r)$ is virtually constant for all x , it will often be advantageous to estimate this variance rather than s_q^2 . This will generally lead to somewhat simpler and more straightforward computations. Any slight bias introduced into the estimates of error will be of little consequence, since it will merely result in a slightly larger or smaller sample being taken.

We frequently require an estimate of the gain in precision due to the use of supplementary information. This is needed in planning a sample survey when a decision has to be reached whether supplementary observations should be taken. It is also required in the planning of the computations in order to decide whether the utilization of available supplementary information is worth the additional computational labour.

In the case of the regression method the relative precision is very simply calculated, since it depends only on the value of the correlation coefficient r , being in fact

$$\frac{1}{1 - r^2}$$

In calculating r due regard must be had to any restrictions imposed by stratification, the same sums of squares and products being used as in the calculation of the regression coefficient and the residual error. The above expression is approximate in that the reduction by 1 of the error degrees of freedom with the regression has been ignored, but this correction will be small relative to errors in the estimation of r .

If an arbitrary value b_0 of the regression coefficient is used the relative precision will be

$$\frac{1}{1 - r^2 + r^2 (1 - b_0/b)^2}$$

The corresponding expression for the ratio method is obtained by writing $\bar{r}$ for b_0 .

Example 8.6.a

From the data of Example 7.17 calculate (a) the number of farms required to give an unbiased estimate of the mean dressing of nitrogen per acre over the farms of the county with a standard error of ± 0.05 cwt., and (b) the number of farms required in each of two equal groups so that the comparison based on the unweighted means of the dressings per acre of the two groups has a standard error of ± 0.05 cwt.

(a) The required number is $67 \times 0.0392^2 / 0.05^2 = 41$. The correction for finite sampling is trivial in this example. Note that either s_q^2 or s_r^2 can be used to arrive at this result.

(b) If the required number in each group is n , the variance of the difference of the means is $2 s_r^2 / n$. Hence $n = 2 \times 0.0541 / 0.05^2 = 43$.

Example 8.6.b

Obtain the expressions for the relative efficiencies given in Example 7.12.b from the above formulæ.

We have $b = 0.6327$, $r = 52,069 / \sqrt{(115,266 \times 82,296)} = 0.537$ and $\bar{r} = 147.36 / 132.08 = 1.116$. Hence the relative precision of the regression

method, compared with the sample plots only, is $1/(1 - 0.537^2) = 1.40$. Similarly the use of differences ($b_0 = 1$) gives a relative precision of $1/\{1 - 0.537^2 + 0.537^2(1 - 1/0.6327)^2\} = 1.23$, the ratio method ($b_0 = 1.116$) gives a value of 1.14, and the regression ($b = 0.55$) gives a value of 1.39. These correspond to the relative efficiencies already tabulated except in the case of the regression, for which we have here neglected the correction for degrees of freedom.

Example 8.6.c

Determine the gains in precision in the estimation of wheat acreages from the random sample of Hertfordshire farms due to the use of supplementary information on acreages of crops and grass, (a) using the ratio method, and (b) using the regression method, without taking account of districts.

The standard errors, already obtained, are ± 7950 for direct estimation without the use of supplementary information (Example 7.2.b), ± 3940 for the ratio method (Example 7.8.a), and ± 4126 for the regression method (Example 7.12.a). The apparent gain in precision due to the ratio method is therefore $7950^2/3940^2 = 4.07$, and that due to the regression method is $7950^2/4126^2 = 3.71$.

The value for the regression appears anomalous, since the formulæ given above indicate that regression may be expected to be at least as efficient (apart from the change in degrees of freedom) as the ratio method. The discrepancy is due to the inclusion of the factor $\bar{x}^2/\bar{x}^2$ in the variance of the ratio estimate. Using the above formulæ with $r = 0.8555$, $b = 0.1932$, $\bar{r} = 0.1522$, we find that the relative precision, compared with direct estimation, is 3.73 for the regression method, and 3.32 for the ratio method. An alternative estimate of the relative precision of the regression and the ratio methods is therefore $3.73/3.32 = 1.12$. This latter value gives a better indication of the average value of the relative precision of the two methods.

8.7 Two-phase sampling

The only case which presents any new features is that in which the first-phase information is used as supplementary information to improve the accuracy of estimates of the second-phase variate y . It has already been pointed out in Section 7.8 that the variance of a two-phase sample is in this case made up of two parts A and B , where

A = variance due to the first-phase sampling, *i.e.* the variance which would be obtained if y were determined for all the units of the first-phase sample,

B = variance due to the second-phase sampling of the first-phase sample (regarded as without error).

To determine B the methods given in Chapter 7 for supplementary information are followed, the effective sampling fraction being n_2/n_1 . To determine A we must use the methods given in the present chapter for the evaluation of the error of a sample of one size and type from the data of a sample of a different size and possibly different type. Thus, if the first-phase sampling is random, and the second-phase sampling is stratified with a variable sampling fraction, it is necessary to calculate the variance of an unstratified random sample of n_1 units from the data of a stratified sample with variable sampling fraction of n_2 units.

Once A and B have been determined the calculation of the relative precision of different possible sampling methods presents no difficulty. If, for example, we wish to ascertain the increase in precision due to taking a two-phase sample of n_1 and n_2 units instead of a single-phase sample of n_2 units, we calculate what the variance A' of a sample of n_2 units would be if the first-phase sampling procedure were followed for a sample of n_2 units. This calculation will follow the same lines as that of A . The relative precision is then $A'/(A + B)$. Similarly the relative precision resulting from the ascertainment of the second-phase information on the n_2 second-phase units only, instead of on all the n_1 units of the sample, will be $A/(A + B)$.

In the simple but general case in which the population is large, and the methods of sampling and estimation are such that the variances of the estimates at each phase are inversely proportional to the numbers of units, apart from the factor $1 - n_2/n_1$, the above relative precisions are capable of simple expression. If the effective variances per unit are s_1^2 and s_2^2 , with $s_2/s_1 = \kappa$ and $n_2/n_1 = \lambda$, we have

$$A = \frac{1}{n_1} s_1^2 \quad A' = \frac{1}{n_2} s_1^2 \quad B = \frac{1}{n_2} \left(1 - \frac{n_2}{n_1}\right) s_2^2$$

Consequently the relative precision giving the gain due to the inclusion of the additional first-phase units is

$$\frac{A'}{A + B} = \frac{1}{(1 - \lambda) \kappa^2 + \lambda}$$

Similarly the loss by not ascertaining the second-phase information over all the first-phase units is given by the relative precision

$$\frac{A}{A + B} = \frac{\lambda}{(1 - \lambda) \kappa^2 + \lambda}$$

Representative values of these fractions are given in Table 8.7. If, for example, the effective standard error per unit is halved by the use of the first-phase supplementary information, $\kappa^2 = \frac{1}{4}$. Consequently, if we introduce two-phase sampling and quadruple the size of the sample for first-phase information only, instead of using single-phase sampling, the amount of information derived from a second-phase unit is increased by a factor of 2.29. Similarly by collecting second-phase information on only $\frac{1}{4}$ of the first-phase

units instead of all the units the amount of information is reduced by a factor of 0.57.

TABLE 8.7—RELATIVE PRECISION OF TWO-PHASE AND SINGLE-PHASE SAMPLING

Two-phase sample :— Single-phase sample :—	n_1 and n_2 n_2			n_1 and n_2 n_1		
κ^2	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{8}$
$\lambda = \frac{1}{2}$	1.33	1.6	1.78	0.67	0.8	0.89
$\lambda = \frac{1}{4}$	1.6	2.29	2.91	0.4	0.57	0.73
$\lambda = \frac{1}{8}$	1.78	2.91	4.27	0.22	0.36	0.53

8.8 Sampling on successive occasions

The relative efficiency of the various estimates can be calculated from the variances given in Section 7.19. When the variances on the different occasions are the same, the relative efficiency of the various estimates, under the conditions set out in Section 6.22, depends only on μ and the correlation r between the successive occasions.

Table 8.8.a gives the efficiencies, relative to those of the overall mean, of the adjusted estimates of the mean on the last occasion (*a*) when there is a sub-sample on the second occasion, and (*b*) with partial replacement, the latter being given for both two and a large number of occasions. Values for $\mu = \frac{1}{2}$ and $\mu = \frac{1}{3}$, and for various values of r , are given. With independent samples or a fixed sample the overall means are fully efficient.

TABLE 8.8.a—SAMPLING ON SUCCESSIVE OCCASIONS : EFFICIENCY, RELATIVE TO THE OVERALL MEAN, OF THE ADJUSTED ESTIMATES OF THE MEAN ON THE LAST OCCASION

r	$\mu = \frac{1}{2}$			$\mu = \frac{1}{3}$		
	Sub-sample	Partial replacement		Sub-sample	Partial replacement	
		Two occasions	Large number		Two occasions	Large number
0	1.00	1.00	1.00	1.00	1.00	1.00
.25	1.03	1.02	1.02	1.02	1.02	1.03
.5	1.14	1.07	1.08	1.09	1.06	1.07
.6	1.22	1.11	1.12	1.14	1.09	1.11
.7	1.32	1.16	1.20	1.20	1.13	1.18
.8	1.47	1.24	1.33	1.27	1.18	1.30
.9	1.68	1.34	1.65	1.37	1.25	1.59
.95	1.82	1.41	2.10	1.43	1.29	2.02
1.0	2.00	1.50	Inf.	1.50	1.33	Inf.

TABLE 8.8.b—SAMPLING ON SUCCESSIVE OCCASIONS : EFFICIENCY, RELATIVE TO THE DIFFERENCE OF THE OVERALL MEANS, OR TO INDEPENDENT SAMPLES (VALUES IN BRACKETS), OF ALTERNATIVE ESTIMATES OF CHANGE

r	$\mu = \frac{1}{2}$		$\mu = \frac{1}{3}$		Fixed sample
	$\bar{y}_h - \bar{y}_{h-1}$	From last two occasions	$\bar{y}_h - \bar{y}_{h-1}$	From last two occasions	
0	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	1.00 (1.00)	(1.00)
.25	1.02 (1.16)	1.02 (1.17)	1.02 (1.22)	1.02 (1.22)	(1.33)
.5	1.10 (1.47)	1.12 (1.50)	1.09 (1.63)	1.11 (1.67)	(2.00)
.6	1.18 (1.69)	1.22 (1.75)	1.15 (1.92)	1.20 (2.00)	(2.50)
.7	1.32 (2.03)	1.41 (2.17)	1.27 (2.37)	1.36 (2.56)	(3.33)
.8	1.60 (2.67)	1.80 (3.00)	1.50 (3.22)	1.71 (3.67)	(5.00)
.9	2.43 (4.41)	3.02 (5.50)	2.21 (5.53)	2.80 (7.00)	(10.00)
.95	3.99 (7.61)	5.51 (10.50)	3.58 (9.76)	5.01 (13.67)	(20.00)

The increase in precision due to the use of partial replacement instead of independent samples or a fixed sample can also be obtained from Table 8.8.a. Thus with a correlation of 0.8 replacement of half the units gives a 24 per cent. increase in precision on the second occasion and a 33 per cent. increase after a number of occasions. With one-third replacement the corresponding percentages are 18 and 30.

Table 8.8.b gives similar efficiencies, relative to the differences of the overall means, or to independent samples (values in brackets), of the estimates of change given by $\bar{y}_h - \bar{y}_{h-1}$ and by the weighted estimate based on the last two occasions only (formula 6.21.b).

In the estimation of change the difference between the overall means of two independent samples is less accurate than the difference of the overall means of a sample with partial replacement. This in its turn is less accurate than the difference between the means of a fixed sample. Thus with a correlation of 0.8 the weighted estimate from the last two occasions, with replacement of half the units, is 3.00 times as efficient as the difference of the means of two independent samples, but only 1.80 times as efficient as the difference of the overall means of the replacement sample. A repeated sample under these circumstances is 5.00 times as precise as a pair of independent samples.

It will be noted that the estimate of change derived from the last two occasions is always somewhat more accurate than the estimate $\bar{y}_h - \bar{y}_{h-1}$. With a correlation of 0.8, for instance, there is a gain in efficiency of 12 per cent. when $\mu = \frac{1}{2}$ and of 14 per cent. when $\mu = \frac{1}{3}$.

Example 8.8

Estimate the relative efficiency of the various estimates of Examples 6.21 and 6.22.

In Example 6.21, $r = 0.847$ and $\mu = \frac{1}{3}$. Consequently, from Table 8.8.a, the relative efficiency of $\bar{y}_w$ and $\bar{y}$ is 1.21. From Table 8.8.b the relative efficiency of the weighted estimate of change and the difference of the overall means is about 2.1. The relative efficiency of the difference of the means of the units common to both occasions and the weighted estimate is given by the weight of the former, namely, 0.929.

The relative efficiency of the estimates of Example 6.22 cannot easily be determined exactly, owing to the variation in the numbers of units from occasion to occasion. With the average value of μ of $\frac{1}{3}$ and a correlation of 0.811, the efficiency of $\bar{y}_h$ relative to the overall mean after a number of occasions will be 1.32 (Table 8.8.a) and that of the estimate $\bar{y}_h - \bar{y}_{h-1}$ of change relative to the difference of the overall means will be about 1.6 (Table 8.8.b).

8.9 Sampling with probability proportional to size of unit

The relative precision of sampling with uniform probability and with probability proportional to size of unit depends on the variance laws to which the material is subject. The case in which the mean r for fixed x is the same for all values of x , and in which the variance of r for fixed x is a function of x , may first be considered.

If the total size of all units is known, we shall be concerned with estimates of $\bar{r}$. If we put $V(x)/\bar{x}^2 = \gamma$ we have the results shown in Table 8.9.a for the three variance laws there given, v being a constant.

TABLE 8.9.a—VARIANCES OF $\bar{r}$

Variance of r for fixed x	Variance of $\bar{r}$	
	Uniform probability	Probability proportional to x
v	$v(1 + \gamma)/n$	v/n
v/x	$v/n\bar{x}$	$v/n\bar{x}$
v/x^2	$v/n\bar{x}^2$	$v(1 + \gamma)/n\bar{x}^2$

In sampling for yield per acre in a crop estimation scheme, for example, the variance of the yield per acre may be expected to be about the same for large and small fields. If in addition there is no marked difference between the mean yields per acre of small and large fields, the precision of sampling

with probability proportional to size relative to sampling with uniform probability will be $1 + \gamma$.

If the mean r for fixed x varies with x the variances of Table 8.9.a will be increased, and the precision of either method, or the relative precision of the two methods, may best be judged by direct analysis of actual data.

If the acreage of the crop has to be determined by the sampling of fields, the relative precision of sampling with probability proportional to size, and with uniform probability, will also depend on the variance of the acreages. The simplest case is that in which the sampling is used to determine which of the fields carry the given crop, and in which the values of $\bar{x}$ and $V(x)$ are the same for the fields of the given crop and for the remaining fields, the number of fields being large. The variance of the proportion p of the total area under the given crop when n' fields are taken is in this case pq/n' with sampling with probability proportional to size, and $pq(1 + \gamma)/n'$ with uniform probability. The relative precision is therefore $1 + \gamma$.

In the case of sampling with probability proportional to size, point sampling will often be used. If the part of the land area which consists of fields cannot be recognized on the map, additional points will have to be visited on the ground, and these must be allowed for in assessing the total number of points required.

In the more complicated cases of sampling with probability proportional to size the same general approach as that adopted in the previous sections must be followed, using the data provided by an actual sample to determine the relevant variances. If the basic data are derived from a sample taken with probability proportional to size, s_r^2 can be calculated from the formulæ of Sections 7.15 or 7.16. The value so obtained may then be used to deduce the size of sample required for a given accuracy.

If the basic data are derived from a sample taken with uniform probability of selection, or if data relating to the whole population are available, the various sizes of unit will occur in proportions which are different from those of a sample taken with probability proportional to size of unit. Consequently a different formula is required for the calculation of s_r^2 . The appropriate formula for a random sample is

$$s_r^2 = \frac{1}{(n-1)\bar{x}} \left[S\left(\frac{y^2}{x}\right) - \frac{\{S(y)\}^2}{S(x)} \right] = \frac{1}{(n-1)\bar{x}} [S(ry) - \bar{r}_u S(y)]$$

where $\bar{r}_u = S(y)/S(x)$. If the individual values of y and r are tabulated the second form of the expression is most convenient for computation.

In the case of a stratified sample the expression within the square brackets must be evaluated for each stratum separately. If the number in each stratum is small and there is no great difference between the $\bar{x}_i$, the separate components can then be aggregated and divided by $(n-t)\bar{x}$. If there are considerable differences between the $\bar{x}_i$ it is best to calculate s_{ri}^2 separately for each stratum, using the separate values in the calculation of $V(Y)$.

Example 8.9.a

From the data of Table 6.19.a construct a frequency distribution of the acreages of sugar-beet fields on old arable land in Norfolk, and hence calculate the relative precision of estimates of the mean yield per acre derived from a random sample of fields taken (a) with probability proportional to size and (b) with uniform probability, on the assumption that the variability of the yield per acre is the same for all sizes of field.

In constructing the frequency distribution, account must be taken of the variable sampling fractions at the two stages of sampling. Since the raising factors at the first stage are nearly proportional to 7, 4, 2, the fields on the small, medium and large farms with a single field of sugar-beet must be counted 7, 4 and 2 times respectively. Similarly a field occurring on a farm with 2 fields of sugar beet must be counted 14, 8 or 4 times, etc.

This procedure gives the frequency distribution shown in Table 8.9.b.

TABLE 8.9.b—FREQUENCY DISTRIBUTION OF THE ACREAGES OF SUGAR-BEET FIELDS

Acreage	Raised No. of fields	Acreage	Raised No. of fields
2	92	12	4
3	43	13	8
4	117	—	—
5	48	20	8
6	54	—	—
7	63	24	6
8	42	—	—
9	4	29	12
10	60	—	—
11	24	48	2
			587

Following the method of Example 7.1.a for grouped data (the acreages being taken as the working units), we find

$$\bar{x} = 6.681, \quad s^2 = V(x) = 30.47, \quad \gamma = 30.47/6.681^2 = 0.681$$

Consequently the relative precision of methods (a) and (b) is 1.68.

Example 8.9.b

From the data of the sample of Hertfordshire parishes taken with uniform probability (Sample A of Section 3.11) estimate the value of s_r^2 for a sample of parishes, stratified by districts, taken with probability proportional to size. Make a similar estimate from the data for all 91 combined parishes.

TABLE 8.9.c—ACREAGES OF CROPS AND GRASS (DIVIDED BY 10), AND OF WHEAT,
IN THE 91 COMBINED HERTFORDSHIRE PARISHES

The parishes selected for samples A and B of Table 3.11.b are indicated by the letters a and b respectively.

x	y	r
363	958	2.6391
227	440	1.9383
250	518	2.0720
289	495	1.7128
242	565	2.3347
1371	2976	2.1707

265

This value is considerably less than the value 0.002649 obtained in Example 7.16. Each estimate, however, is based on only 10 degrees of freedom, so that the discrepancy is not exceptionally large. The corresponding value from the data for all 91 combined parishes, calculated in the same manner, is 0.002222. This calculation is left as an exercise for the reader.

Example 8.9.c

Compare the relative precision, in the estimation of wheat acreage, of samples of Hertfordshire parishes taken with uniform probability and with probability proportional to size, by calculating the expected standard errors of samples of types *A* and *B* of Table 3.11.b.

The data for all 91 combined parishes give a value of s_d^2 of 23,483 when districts are eliminated and the same ratio is taken for all districts, and a value of 22,427 when different ratios are taken for the different districts.

In calculating the expected standard error the formula of Section 7.10 may be used, so as to allow for the variation in sampling fraction from district to district. The factors $\mathbf{X}_i^2/\{S_i(x)\}^2$ may be replaced by $1/f_i^2$ since we are considering the average error to be expected over a series of similar samples. This will lead to a slight underestimation of the average error.

We find $\Sigma (1 - f_i) n_i / f_i^2 = 402.83$, and consequently $V(Y) = 9.460 \times 10^6$ when the same ratio is taken for all districts, and 9.034×10^6 when different ratios are taken.

Similarly, in the case of sampling with probability proportional to size, from the results already given in Example 7.16, and the value of s_r^2 given in Example 8.9.b, we find $V(Y) = 8.263 \times 10^6$.

The standard errors corresponding to these variances have already been given in Table 3.11.b.

The relative precision of sampling with probability proportional to size, and with uniform probability using a single value of the ratio, is therefore $9.460/8.263 = 1.14$. There is thus a gain in precision of 14 per cent., but it must be recognized that sampling with probability proportional to size will result in parishes of larger average size being included in the sample. Neglecting the disturbance due to the probability being only approximately proportional to size, the average size of parish in this case will be given by $S(x^2)/S(x)$, where the summations are taken over the whole population (or a sample selected with uniform probability). This gives an average size of 3244 acres of crops and grass, compared with the arithmetic mean of 3000 acres, *i.e.* an average size greater by 8 per cent.

8.10 Interpretation of the analysis of variance

The analysis of variance can be interpreted in the manner set out below. This interpretation is of particular use when we are concerned with multi-stage sampling, and with the effect of change of size of the sampling units.

If the units fall into groups of any kind, such as strata, the unit values of a variate y can be regarded as made up of the sum of two parts, one, u , which varies from group to group but has a fixed value for all units of a particular group, and the other, v , which varies from unit to unit independently of the groups. The variances of u and v may be denoted by U and V respectively. Thus u and v may be random sample values from normal distributions, though the condition of normality is not necessary. In this hypothetical framework zero mean can be assigned to the parent distribution of v without loss of generality, but even so the mean of the v 's for all the units of a finite population, or for all the units of a particular group, will not be exactly zero, and consequently the group means are not exactly equal to the u 's. For this reason the values of u and v cannot be uniquely determined from the values of y .

The mean squares of the analysis of variance provide estimates of U and V . If A and B are the mean squares between and within groups, C is the overall mean square, k is the number of units in each group, and h the number of groups, we have

$$\begin{aligned} A &= kU + V \\ B &= V \end{aligned}$$

Hence

$$U = (A - B)/k$$

We also have, from the analysis of variance, $(hk - 1)C = h(k - 1)B + (h - 1)A$. Consequently if σ^2 is the overall variance and σ_1^2 the variance within groups we have, from formula 8.3,

$$\begin{aligned} s^2 &= U(h - 1)/h + V \\ s_1^2 &= V \end{aligned}$$

The factor $(h - 1)/h$ is analogous to the correction for sampling from a finite population.

The relative precision of stratified and random sampling will be obtained by taking the groups as strata. We then have, with t strata,

$$\frac{s^2}{s_1^2} = 1 + \frac{t - 1}{t} \cdot \frac{U}{V}$$

An alternative formulation is possible in terms of the *intra-class correlation*, i.e. the correlation between members of the same stratum when the strata themselves are regarded as a random sample from an infinite set of similar strata (R. A. Fisher, *Statistical Methods for Research Workers*, Section 40). The estimate r_i of this correlation is given by

$$r_i = \frac{A - B}{A + (k - 1)B} = \frac{U}{U + V}$$

and consequently

$$\frac{s^2}{s_1^2} = 1 + \frac{t - 1}{t} \cdot \frac{r_i}{1 - r_i}$$

Looked at from this point of view, the intra-class correlation coefficient may be regarded as a quantitative expression of association which is alternative

to the ratio $\mathbf{U}/\mathbf{V}$. In this book we shall use the concept of additive components of variance, since this appears to be more easily capable of generalization, and is otherwise preferable to the concept of intra-class correlation.

When there is compensation between the different units of the same stratum the definition of $\mathbf{U}$ as a variance breaks down, and has to be extended (see Yates and Zaccopani, 1935, H). Complete compensation occurs when all the strata means (or the first-stage units in a two-stage scheme) are equal. In this case $\mathbf{K}\mathbf{U} + \mathbf{V} = 0$, *i.e.* $\mathbf{U} = -\mathbf{V}/\mathbf{K}$, where $\mathbf{K}$ is the number of units in each group of the population. Negative values of $\mathbf{U}$ between 0 and $-\mathbf{V}/\mathbf{K}$ are therefore admissible.

8.11 Multi-stage sampling

The sampling variance of two-stage sampling can be divided into two parts, A and B , where

A = variance due to the first-stage sampling when there is complete ascertainment at the second stage, *i.e.* when all the second-stage units which go to make up the selected first-stage units are known,

B = variance due to the second-stage sampling of the selected first-stage units.

Thus the formula of Section 7.17 for $V(\bar{y})$ in two-stage random sampling may be rewritten

$$V(\bar{y}) = \frac{1-f'}{n'} s_0'^2 + \frac{1-f''}{n' n''} s''^2 \quad (8.11.a)$$

where

$$s_0'^2 = s'^2 - \frac{1-f''}{n''} s''^2 \quad (8.11.b)$$

The first term constitutes part A and the second part B .

The second term will be recognized as $(1-f'')(1-f)$ times the variance that would be obtained with single-stage sampling of the second-stage units, the same total number of second-stage units being taken, with the first-stage units as strata and uniform sampling fraction f . If f'' is small, therefore, the first term gives the increase in variance due to the adoption of the two-stage process.

The above subdivision is alternative to that given in Section 7.17. Part A is dependent only on the first-stage sampling, being unaffected by the intensity or type of sampling at the second stage. This fact considerably simplifies the problem of determining the sampling errors for different intensities of sampling at the two stages: with the subdivision of Section 7.17 the variation in s'^2 for different intensities of sampling at the second stage has to be taken into account.

The only new point that arises in the estimation of the relevant variances is the determination of part A from the data of a two-stage sample. In general this simply requires that the variance per first-stage unit due to the second-stage

sampling of the first-stage units be deducted from the variance per first-stage unit calculated from the sample. Thus for a two-stage random sample formula 8.11.b is used.

It is often helpful to carry out an analysis of variance on data derived from a two-stage sampling process. The situation is simplest when the number of sampled second-stage units n'' in each first-stage unit is the same. Each stage of the analysis then follows the same pattern as the analysis of a single-stage sample of the same type. At the first stage, however, the values entering into the analysis must be either the means or the totals of the second-stage unit values. It is customary (though not essential) to tabulate the sums of squares of the first stage in terms of the second-stage units. If the first-stage unit means are used, therefore, all sums of squares at the first stage must be multiplied by n'' , while with totals all sums of squares must be divided by n'' .

In the case of two-stage random sampling, for example, the degrees of freedom and mean squares will be

		Degrees of freedom	Mean square
Between first-stage units		$n' - 1$	$n'' s'^2 = V + n'' U$
Within first-stage units	between		
second-stage units		$n' (n'' - 1)$	$s''^2 = V$
TOTAL		$n' n'' - 1$	

We then have

$$V(\bar{y}) = \frac{1-f'}{n'} U + \frac{1-f}{n} V \quad (8.11.c)$$

where n is the total number of second-stage units and f is the overall sampling fraction ($n = n' n''$ and $f = f' f''$). The second term of this subdivision is the estimate of the variance that would be obtained with single-stage sampling of the second-stage units, the same total number of sampling units being taken, with first-stage units as strata, and uniform sampling fraction. The analysis of variance therefore provides a further alternative subdivision of the sampling variance.

The results are similar with stratification with uniform sampling fraction at either or both stages.

When one or both the sampling fractions are variable, or when the numbers of second-stage units in the different first-stage units are unequal, the analysis of variance becomes more complicated and the direct approach is often simplest. With moderate inequality in the n'' the analysis of variance of the first-stage units may be carried out on the *means*, with multiplication of the mean squares by $\bar{n}''$, or better by the harmonic mean of the n'' , *i.e.* the reciprocal of the mean of the reciprocals.

Alternatively the whole analysis may be carried out in terms of the second-stage units. In this case both the means (in terms of the second-stage units) and totals of the first-stage units are tabulated, the sums of squares being obtained

by the "mean $\times$ total" rule, *i.e.* every mean is multiplied by the corresponding total.

If the second method is used n'' can be replaced by $\bar{n}''$ in the expression $n''s'^2 = V + n''U$ given above for the mean square for first-stage units of a random sample, or better by $\bar{n}_0''$ where

$$\bar{n}_0'' = \{S(n'') - S(n''^2)/S(n'')\}/(n' - 1) \quad (8.11.d)$$

With a stratified sample a value for $\bar{n}_0''$ is calculated for each stratum and a weighted mean taken, weighting by the degrees of freedom contributed to the within-strata sum of squares (Cochran, 1939, A).

These alternative methods of analysis are not exactly equivalent, but we cannot discuss their differences here, beyond stating that the first method is generally best when all the first-stage units are of approximately the same size and the variation in the numbers of second-stage units per first-stage unit is due to extraneous causes, whereas the second method is likely to be preferable when the first-stage units vary greatly in size and the number of second-stage units per first-stage unit is about proportional to this size.

The above methods can easily be extended to multi-stage sampling with more than two stages.

Example 8.11.a

Calculate the expected sampling errors of the wheat acreages derived from the two-stage sample B_1 of Hertfordshire farms of Table 3.11.b, and discuss the effects of varying the number of parishes in the sample, with adjustment of the second-stage sampling fraction so as to give the same total number of farms in the sample.

Part *A* of the variance has already been determined in Example 8.9.c. We have $A = 8.263 \times 10^6$.

The determination of part *B* requires the evaluation of the variance of the r for individual parishes due to the second-stage sampling of these parishes. These variances were evaluated separately for each of the 17 parishes of the sample, using method (a) of Section 7.9. The mean value of these variances $V''(r)$ was found to be 0.003575.

The equation of estimation of the total acreage is $Y = \sum \mathbf{X}_i \bar{r}_i$. The second-stage variance of $\bar{r}_i$ is $V''(r)/n_i$, and part *B* of the variance is therefore given by

$$B = V''(r) \sum \mathbf{X}_i^2/n_i = 0.003575 \times 45.429 \times 10^8 = 16.24 \times 10^6$$

Hence $V(Y) = 24.50 \times 10^6$.

Exact treatment of the effects of varying the number of parishes is complicated by the fact that the first-stage sampling fractions are bound to vary somewhat from district to district, and that the number of farms per parish is also variable.

EFFICIENCY

Ignoring these sources of disturbance we may write for any number n' of parishes and n'' of farms per parish

$$V(Y) = \frac{1-f'}{n'} \alpha + \frac{1-f''}{n' n''} \beta$$

The values of α and β can be determined from the values of A and B . The mean number of farms per parish is $N'' = 2496/91 = 27.429$. Putting $n' = 17$, $f' = 17/91$, $f'' = \frac{1}{4}$, and $n'' = \frac{1}{4} \times 27.429 = 6.857$, we have

$$\alpha = \frac{17}{1 - 17/91} \times 8.263 \times 10^6 = 172.7 \times 10^6$$

$$\beta = \frac{17 \times 6.857}{1 - \frac{1}{4}} \times 16.24 \times 10^6 = 2524.2 \times 10^6$$

The effect of any variation in n' and n'' can now be determined from the formula for $V(Y)$. If the total number of farms $n (= n' n'')$ is to be kept fixed, the formula is best rewritten in the form

$$V(Y) = (\alpha - \beta/N'')/n' + \beta/n - \alpha/N' \\ = \{80.71/n' + 2524.2/n - 1.8982\} \times 10^6$$

with the checks that, when $n' = 91$ and $n = 2496$, $V(Y)$ is zero, and when $n' = 17$ and $n = 17 \times 6.857 = 116.57$, it equals the value given above.

The values of $V(Y)/10^6$ for 5, 10, 20, and 30 parishes and a number of farms, 116.57, approximately the same as that of the actual sample, are 35.9, 27.8, 23.8 and 22.4 respectively. If all 91 parishes were sampled the corresponding variance would be 20.6. There is thus no great gain in taking more than 20 parishes when sampling within parishes is with a uniform sampling fraction.

It should be noted that the use of the above formula when the fraction of parishes sampled is large is unrealistic, in that sampling with probability proportional to size could not be adopted in such cases. It serves, however, to illustrate the use of the similar formulæ which could be developed for sampling with uniform probability at the first stage.

Example 8.11.b

Repeat the analysis of Example 8.11.a for the sample B_2 of Table 3.11.b.

Part A of the total variance will be the same as in sample B_1 .

Part B can be calculated in the same manner as in Example 8.11.a, using the method of Section 7.10, with a separate value of the ratio for each parish. This gives $V''(r) = 0.0008167$, and $B = 3.710 \times 10^6$. Hence $A + B = 11.97 \times 10^6$.

The expression of $V(Y)$ in terms of n' and n'' is complicated by the variable sampling fraction at the second stage. As a first approximation we may take an average sampling fraction f'' for the given sample, and use this to obtain

a formula of the same form as in Example 8.11.a. This gives $f'' = 0.29121$ and we then find

$$V(Y) = \{146.83/n' + 710.76/n - 1.8982\} \times 10^6$$

with checks as before.

This formula gives values of $V(Y)/10^6$ for 5, 10, 20 and 30 parishes and 135.8 farms of 32.7, 18.0, 10.7 and 8.2 respectively. With more accurate sampling of the farms within the selected parishes, therefore, there is a more marked decrease in variance as the number of parishes is increased. The above values underestimate the decrease, as with the reduction in the number of farms per parish the change in the second-stage sampling fractions will result in a somewhat smaller increase in the variance than that given by the formula. More accurate values could be obtained by recalculating the second-stage variances with various intensities of second-stage sampling, using graphical methods for interpolation between the calculated values.

Example 8.11.c

Investigate the relative precision of the determination of the acreages of crops by the measurement of the areas of fields and part fields included in a sample of rectangular areas, and the use of grids of points covering these areas (Section 4.24).

If a sample area has an area a and the proportion of the area occupied by a given crop is p , the area y occupied by this crop in this sample area equals ap . If a random set of n points is taken over the area the variance of the estimate y given by the proportion of points falling in the given crop is

$$V(y) = pq a^2/n$$

This variance will be additional to the sampling variance of the y over the sample areas. This latter variance depends on the sampling method and the variability of the y from area to area, and can only be determined from actual sample data.

As an example we may consider the case in which the areas are randomly selected, and the frequency distribution of the areas with proportions 0.0, 0.1, . . . of the given crop is as follows:

Proportion, p	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Frequency, ϕ	0.05	0.15	0.20	0.15	0.12	0.10	0.07	0.05	0.03	0.03	0.05

The average variance due to the point sampling is given by

$$\Sigma \phi V(y) = (a^2/n) \Sigma \phi p q = 0.1656 a^2/n$$

The sampling variance of the y is given by

$$V(y) = \Sigma \phi p^2 a^2 - (\Sigma \phi p a)^2 = 0.069024 a^2$$

The proportional increase in variance due to point sampling is therefore

$$0.1656/0.069024 n = 2.40/n$$

With 9 points there is an increase of 27 per cent. in the variance and with 16 points an increase of 15 per cent. In the latter case about one-seventh more areas will be required for the same accuracy. Against this must be set the fact that only fields in which the points fall need be examined and recorded. The occurrence of mixed crops and systematic location of the points on a rectangular grid will also reduce the sampling variance.

8.12 An example of a pilot sampling scheme for crop estimation

In order to investigate the practicability of obtaining estimates of the yields of cereal crops in the United Kingdom by the harvesting of sample areas, the yields of a number of wheat fields were determined by this method in each of the years 1934-1938 (Cochran, 1939, A). Fields were taken in several districts each year, one or two fields being selected at random from the fields growing wheat on each chosen farm. The selection of farms in each district was not random, the farms being taken in the neighbourhood of the centres at which the investigators were located.

The sampling of the individual fields followed the lines described in Section 4.29, the fields being traversed in the direction of the rows, along two lines selected at random. Two sets of unit areas were taken from each line. Each unit area consisted of $\frac{1}{4}$ metre of each of 6 contiguous rows. For the most part, sets each contained three unit areas, equally spaced along the line, with a random starting point.

The yields of grain obtained in 1937 are shown in Table 8.12.a. The mean yield of all the unit areas in each set is given. In order to allow for differences in row spacing on the different fields the yields have been reduced to a 6-inch row spacing, and therefore represent the yields in grams of areas of $\frac{1}{4}$ metre $\times$ 3 ft. Fields on the same farm are indicated by brackets. In District III, where three fields from a single farm were sampled, each field was growing two varieties which were sampled separately.

The analysis of variance was carried out in units of the totals of the four sets, *i.e.* on yields of areas of 1 metre $\times$ 3 ft. or 0.000226 acres. Thus the sum of squares of the sets is multiplied by 4, and the sum of squares of the line totals by 2.

The sums of squares for 1937 can be obtained from Table 8.12.a by calculating the sum of squares for each classification, disregarding the others, and deducting the sum of squares corresponding to the next higher classification. The rule of "mean $\times$ total" or "total²/(number of units)" is followed in each case. Thus the correction for the mean is $11,706^2/39 = 3,513,601$. The sum of squares for districts is

$$\frac{1}{4}(1098)^2 + \frac{1}{3}(909)^2 + \dots - 3,513,601 = 54,224$$

The sum of squares for farms is

$$\begin{aligned} \frac{1}{2}(422)^2 + \frac{1}{2}(676)^2 + \frac{1}{2}(619)^2 + 290^2 + \frac{1}{8}(2053)^2 + \dots \\ - 3,513,601 - 54,224 = 132,062 \end{aligned}$$

The arithmetical work can be simplified by omitting items which are repeated in more than one sum of squares. In particular the sum of squares for varieties is

$$364^2 + 405^2 + 287^2 + \dots - \frac{1}{2}(769)^2 - \frac{1}{2}(597)^2 - \frac{1}{2}(687)^2 = 1326$$

TABLE 8.12.a—SAMPLING OF WHEAT FIELDS, 1937 : MEAN YIELDS OF GRAIN PER UNIT AREA (0.0000565 ACRES) IN GRAMS

		District I				District II			District III						
1st line	Set 1	47	48	75	105	93	58	76	92	89	89	75	70	80	
	Set 2	63	51	71	82	84	78	57	83	111	58	72	85	97	
2nd line	Set 3	67	45	75	97	75	68	79	93	90	70	82	76	111	
	Set 4	55	46	85	86	80	83	78	96	115	70	81	102	66	
		232	190	306	370	332	287	290	364	405	287	310	333	354	
		422	676			619			769	597			687		
Totals		1098				909			2053						
		District IV												District V	
1st line	Set 1	29	45	57	69	78	59	68	97	60	65	81	77	60	
	Set 2	21	39	63	55	109	59	56	88	53	59	94	74	88	
2nd line	Set 3	29	69	46	21	90	58	74	109	43	49	93	44	84	
	Set 4	31	57	66	40	51	53	61	95	48	71	92	57	97	
		110	210	232	185	328	229	259	389	204	244	360	252	329	
		320	417											581	
Totals		2750												581	
		District VI													
1st line	Set 1	66	93	55	127	84	80	81	93	21	84	87	79	90	
	Set 2	73	70	56	106	80	86	107	106	63	51	67	79	117	
2nd line	Set 3	64	80	83	84	63	88	135	71	50	82	135	71	112	
	Set 4	73	67	60	98	89	110	82	83	29	80	114	89	122	
		276	310	254	415	316	364	405	353	163	297	403	318	441	
		586	669			680									
Total		4315													

Furthermore the sum of squares corresponding to the difference between any two totals containing the same number of units can be obtained by squaring

the difference and dividing by twice the number of units in either total. Thus the sum of squares due to varieties is also given by $\frac{1}{2}(41^2 + 23^2 + 21^2)$. This gives a useful check in cases in which, as here, many of the sums of squares depend on differences of pairs of values. Thus the sum of squares between sets is given by $2(16^2 + 12^2 + 3^2 + 1^2 + \dots) = 52,368$, that between lines by $12^2 + 8^2 + \dots = 33,924$.

The sum of squares between fields within farms (excluding District III) is given by $\frac{1}{3}(42^2 + 64^2 + 45^2 + 100^2 + \dots) = 27,702$. The sum of squares between fields within farms for District III has to be calculated in the ordinary manner, since there are three fields, and is $\frac{1}{2}(769)^2 + \frac{1}{2}(597)^2 + \frac{1}{2}(687)^2 - \frac{1}{6}(2053)^2 = 7,401$, giving a corresponding total sum of squares of 35,103.

Those not fully familiar with the analysis of variance technique should recalculate the sums of squares of this example in the various alternative ways indicated.

For general purposes it is best to convert the mean squares into some common units such as (cwt. per acre)². The conversion factor is here 0.0075861. This is done in Table 8.12.b, which also shows the results of the similar analyses for the other four years of the investigation.

TABLE 8.12.b—ANALYSIS OF VARIANCE PER FIELD OF YIELDS OF WHEAT GRAIN (CWT. PER ACRE)

	1934		1935		1936		1937		1938	
	d.f.	m.s.	d.f.	m.s.	d.f.	m.s.	d.f.	m.s.	d.f.	m.s.
Between districts .	4	66.5	6	318.4	4	79.4	5	82.3	4	206.8
Within districts between farms .	11	38.9	12	27.1	7	62.2	19	52.7	14	65.3
Within farms between fields .	—	—	15	22.8	8	31.2	11	24.2	8	12.1
Within fields:										
Sampling error .	16	5.33	40	6.20	22	11.39	39	6.60	28	9.80
Between sets .	32	2.11	80	2.18	45	2.52	78	5.09	55	4.78
Mean yield .	29.1		23.3		24.3		26.2		30.7	

The mean squares for the same component of variance in the different years are not estimates of precisely the same quantities, owing to variation in the numbers of fields per farm, etc. In view of the small number of degrees of freedom in each year, however, we shall not lose much information by pooling all the years, weighting the mean squares in proportion to the degrees of freedom. This pooled estimate is shown in Table 8.12.c.

From the degrees of freedom we may deduce that there are 91 farms and 133 fields in all in the sample, *i.e.* a mean of 1.46 fields per farm. Denoting the variance per set by V_1 , the additional components of variance per line,

per field and per farm by V_2 , U_1 and U_2 respectively, and ignoring the fact that a few fields have more than two lines and that the number of fields per farm is variable, we have the mean square equivalences shown in Table 8.12.c. Hence $V_1 = 14.0$, $V_2 = 8.4$, $U_1 = 15.0$, $U_2 = 18.2$.

TABLE 8.12.c—COMBINED ANALYSIS OF VARIANCE

	Degrees of freedom	Mean square	Estimate
Between districts	23	162.3	
Within districts between farms	63	49.3	$\frac{1}{4} V_1 + \frac{1}{2} V_2 + U_1 + 1.46 U_2$
Within farms between fields	42	22.7	$\frac{1}{4} V_1 + \frac{1}{2} V_2 + U_1$
Within fields between lines	145	7.69	$\frac{1}{4} V_1 + \frac{1}{2} V_2$
Within lines between sets	290	3.50	$\frac{1}{4} V_1$

The value of U_2 is an underestimate, firstly because we have used the mean number of fields per farm, instead of calculating the correct value of $\bar{n}_0''$ from formula 8.11.d, and secondly because of the fact that the sample of farms was not random. For 1937 the value of $\bar{n}_0''$ is 1.29, compared with the value of $\bar{n}''$ of 1.44.

From the above estimates of the different components of variance we may calculate the variance to be expected with a sample of any given type and size. If the unweighted mean of the yields per acre of the different fields can be taken as the estimate of the mean yield per acre over the country, *i.e.* if the potential bias due to association of yield per acre with size of field, etc., can be ignored, the variance of the mean yield per acre with a fixed amount of sampling of individual fields and with equal numbers of fields taken from all selected farms will depend solely on the number of farms and the number of fields in the sample. If these are n_1 and n_2 respectively the variance of the mean yield per acre with the same amount of sampling per field as that actually adopted will be

$$U_1'/n_2 + U_2/n_1$$

where $U_1' = \frac{1}{4} V_1 + \frac{1}{2} V_2 + U_1 = 22.7$. Thus with 200 fields from 100 farms, 2 fields per farm, the variance with the above values of the components of variance is 0.296. This is equivalent to a standard error of 0.54 cwt. per acre, or 2.0 per cent. of the mean yield.

This is an over-simplification of the practical situation. In general the possibility of bias cannot be ignored, and a properly weighted mean must therefore be taken. Any statement in general terms would be difficult, since

the weighting depends on the variation in numbers and acreages of the fields on the individual farms and the sampling method adopted. Given the numbers and acreages of the fields on an adequate sample of farms, however, the weighting coefficients for any chosen method of sampling can be determined. If these are denoted by w , and if $[w]$ denotes the sum of the weights for all the sampled fields on a farm, the variance of the weighted mean will be

$$\{U_1' S(w^2) + U_2 S([w]^2)\} / \{S(w)\}^2$$

Thus the relative precision of alternative methods of sampling can be evaluated without difficulty.

The above procedure is approximate in another respect which is not entirely irrelevant to the practical situation. It has been assumed that the component of variance from field to field on the same farm, and that between farms, are independent of the size of the farm. This is not likely to be strictly true, and may introduce appreciable inaccuracy when a variable sampling fraction is used for farms of different sizes.

It will be noted that no corrections for sampling from a finite population are necessary, provided the fraction of the farms in the sample is small. The second-stage sampling fraction of fields from farms may be large, but this fraction does not enter into the formula for the partition of the variance given by the analysis of variance, as for example is shown by formula 8.11.c.

Although the variance of the unbiased estimate depends on the acreages, and therefore cannot be easily formulated in general terms, certain general statements about the relative precision of different types of sample can be made from the above results.

In the first place we may consider the effect of varying the amount of sampling of the selected fields. If the number of sets per line is reduced to one, for example, the sampling variance per field will be $\frac{1}{2} V_1 + \frac{1}{2} V_2 = 11.2$, instead of 7.7. If at the same time the number of lines is increased to four, the sampling variance per field will be $\frac{1}{4} V_1 + \frac{1}{4} V_2 = 5.6$.

There is little to be gained by increase in the accuracy of the determination of the yields of individual fields, however. With one field per farm the effective variance per field if the yields are determined without error will be $U_1 + U_2 = 33.2$, instead of 40.9. Consequently with the intensity of sampling actually adopted the relative accuracy is 0.81. Doubling the number of lines per field with two sets per line would only increase the accuracy by 10 per cent.

The question of whether to sample one or more fields per farm requires more consideration. For farms growing a given number of fields of wheat (greater than one) the variance if two fields are sampled will be $\frac{1}{2} U_1' + U_2 = 29.6$, whereas if one field is sampled the variance will be 40.9. The whole question of the methods of sampling farms and fields is bound up with the question of costs, and will be further considered in Section 8.17.

The effect of varying the number of unit areas per set cannot be precisely determined from the above analysis. If the unit areas of each set were randomly and independently located, the variance of the set means would be inversely

proportional to the number of units per set, and would be determined from the variance between sets within lines. With the even spacing of units within the set, however, we may expect the reduction in variance with increasing numbers of unit areas to be somewhat greater than in the case of random location. From the basic data giving the yields of the separate unit areas it would be possible to determine the variance per unit area within sets, and from this variance and the variance between sets within lines a fair idea of the departure from the random law could be obtained.

As a first approximation, however, we may assume that the random law holds. In this case, taking two instead of three unit areas per set would multiply the value of $\bar{V}_1$ by $3/2$, and would therefore raise the value of $U_1' + U_2$ from 40.9 to 42.6. It was in fact recognized after the first year's work that there was little to be gained from having more than a small number of unit areas per set, and the number, which was five in the first year, was then reduced to three.

8.13 A special case of two-stage sampling

The possibility of sampling from within strata with probability proportional to size at the first stage, and with second-stage sampling fractions so chosen that the overall sampling fraction is uniform, has already been mentioned in Section 3.10 and subsequently.

This case is of considerable practical importance, and also provides a useful example of the application of the above methods to the more complicated types of two-stage sampling.

From the results already given in Sections 7.16 and 7.17 we have

$$V(Y) = \sum s_i'^2 \mathbf{X}_i^2 (1 - f_i')/n_i' + \sum f_i' \mathbf{X}_i^2 V''(\bar{r}_i) \quad (8.13)$$

where $V''(\bar{r}_i)$ is the estimated second-stage variance of $\bar{r}_i$.

We will consider the case in which the sampling at the second stage is random (or stratified with uniform sampling fraction) and the number of second-stage units is taken as the measure of size. If n_i' first-stage units are selected from the i th stratum the probability of selection of the j th unit will be $n_i' \mathbf{N}_{ij}/\mathbf{N}_i$, where $\mathbf{N}_{ij}$ is the number of second-stage units in the j th first-stage unit, etc. The second-stage sampling fraction for this unit, if selected, will be $f \mathbf{N}_i/n_i' \mathbf{N}_{ij}$, where f is the uniform overall sampling fraction. The number of second-stage units selected will be $f \mathbf{N}_i/n_i'$. Thus the same number of second-stage units will be selected from each of the selected first-stage units in a given stratum. If the variance $\sigma_i''^2$ of y per unit at the second stage can be taken as constant for the whole of the i th stratum the estimated variance of r_{ij} ($= \bar{y}_{ij}$) will be

$$V''(r_{ij}) = s_i''^2 \frac{(1 - f_{ij}'')}{f \mathbf{N}_i/n_i'}$$

which is constant for all the selected units of the i th stratum except for the

factor $(1 - f_{ij}'')$. Provided all the f_{ij}'' are moderately small it will be sufficient to replace them by $f_i'' = f/f_i'$. We then have

$$V''(\bar{r}_i) = s_i''^2 (1 - f_i'')/f N_i$$

The $\mathbf{X}_i$ of formula 8.13 will be replaced by $\mathbf{N}_i$, and some form of pooling can be adopted to estimate an average value $s_r'^2$ of $s_{ri}'^2$. In cases in which the $s_{ri}'^2$ are likely to vary markedly, weights corresponding to those given by the first term should be used.

We then have

$$V(Y) = s_r'^2 \sum N_i^2 (1 - f_i')/n_i' + \sum f_i' s_i''^2 N_i (1 - f_i'')/f$$

Following the previous procedure this may be re-written as

$$V(Y) = s_{ro}'^2 \sum N_i^2 (1 - f_i')/n_i' + \sum s_i''^2 N_i (1 - f_i'')/f$$

where

$$s_{ro}'^2 = s_r'^2 - \frac{\sum (1 - f_i') s_i''^2 N_i (1 - f_i'')}{f \sum N_i^2 (1 - f_i')/n_i'}$$

If the f_i' are approximately equal, as will usually be the case, and the $s_i''^2$ are the same for all strata, we have

$$V(Y) = s_{ro}'^2 (1 - f') \sum N_i^2/n_i' + s''^2 N (1 - f'')/f$$

$$s_{ro}'^2 = s_r'^2 - \frac{s''^2 N (1 - f'')}{f \sum N_i^2/n_i'}$$

The second term of $V(Y)$ will be recognized as $(1 - f'')(1 - f)$ times the variance which would be obtained with single-stage sampling of the second-stage units with uniform sampling fraction and the first-stage units as strata. If f'' is small, therefore, the first term gives the increase in variance due to the adoption of the two-stage process, as in the case of two-stage random sampling.

8.14 Effect of change in size of the sampling units

If the population is divided into N large units, each of which is subdivided into K small units, and if a sample of k small units from each of n large units is taken, an analysis of variance between and within large units can be made, and the components of variance U and V estimated as in Section 8.10.

The estimate of the overall variance between small units will be given, as before, by $V + U(N - 1)/N$. That between large-unit means of k small units will be given by $1/k$ times the expectation of the mean square between large units, i.e. by $V/k + U$. The variance between large-unit means when all K small units of each large unit are included will be $V/K + U$.

These results enable us to determine the effect on the sampling error of the alternatives of using the large or the small units as sampling units. It will be noted that for this determination it is not necessary to have data in which all the small units that go to make up the selected large units are observed. The analogy with two-stage sampling will be apparent.

Various extensions of these results are of interest. If the large units are stratified, with N_t units per stratum in the population and n_t in the sample, there will be a further between-strata component of variance U_t , and the mean squares in the analysis of variance will provide estimates as follows:—

Between strata	..	..	$V + k U + k n_t U_t$
Within strata between large units	..	..	$V + k U$
Within large units between small units	..	..	V

The estimates of the different variances will then be:

Small units within strata	..	..	$V + U (N_t - 1)/N_t$
Large units (means) within strata	..	..	$V/K + U$
Large units (means) overall	..	..	$V/K + U + U_t (t - 1)/t$

The first two variances are the same as previously, except that N is replaced by N_t .

The above approach enables us to determine the effect of simultaneous change of size of unit and size of strata. This is relevant when the strata can be of any size, and the size is therefore chosen to contain two units (or one unit) per stratum. If the size of unit is halved and the amount of material in the sample remains the same, for example, there will be twice as many units, and the size of the strata can therefore be halved. We shall then require a four-fold analysis of variance into whole strata, half-strata, whole units, and half-units. The minimum amount of data required for this purpose will be two whole units (of which each half-unit is separately recorded) in each half-stratum.

The expressions for mean squares and variances will be similar to those given above, N_t being the number of whole units per half-stratum in the population, and t the number (two) of half-strata per stratum. These expressions are as follows:—

Mean squares:

Within whole strata between half-strata	..	..	$V + 2 U + 4 U_t$
Within half-strata between whole units	..	..	$V + 2 U$
Within whole units between half-units	..	..	V

Variances:

Half-units within half-strata	..	..	$V + U (N_t - 1)/N_t$
Whole units within half-strata	..	..	$\frac{1}{2} V + U$
Whole units within whole strata	..	..	$\frac{1}{2} V + U + \frac{1}{2} U_t$

Since there will be half as many whole units as half-units the relative precision of the two methods of sampling will be given by

$$\frac{V + 2 U + U_t}{V + U (N_t - 1)/N_t}$$

8.15 Variation in size of strata

When the strata boundaries are arbitrary, the size of the strata may be varied in such a manner that a fixed number of units require to be selected

from each stratum, whatever the size of the sample. The strata will naturally be taken as small as possible, *i.e.* so as to contain two units if a rigorous estimate of error is required, or one unit otherwise.

In order to determine the size of sample required for a given accuracy under these conditions it is necessary to know the relation between the size of the strata and the within-strata variance. The simplest way in which this relation can be determined is to obtain data for all units of a representative sample of the largest strata that are of interest. Strata of any smaller size can then be constructed, and the within-strata variances calculated.

A minor difficulty in this construction is that the original strata will only be exactly subdivisible into strata of smaller size if these contain numbers of units which are integral fractions of the numbers in the original strata. If in area sampling the strata are also to be of the same shape, only squares of integral fractions, *i.e.* $\frac{1}{4}$, $\frac{1}{9}$, . . . , will give exact subdivision. For strata of intermediate size there will therefore be a certain amount of arbitrariness in the location of the strata boundaries. Some objective rule must therefore be followed. If it appears desirable, overlapping strata may be used. Thus in a case in which the data cover a set of isolated squares, four sets of smaller squares may be taken within each large square, each set having a corner point coincident with one corner of the large square.

If the smallest strata likely to be of interest each contain a large number of units, the collection of data in full for all the units of these basic strata is likely to be laborious. Instead a random sample of such units may be taken. In this case the within-strata variances can be estimated by means of an analysis of variance similar to that used for change in size of sampling units (Section 8.14). If the small units of that section are taken as equivalent to the units of the present case, the large units as equivalent to the basic strata, and the strata as equivalent to the larger strata, the same expressions hold.

This procedure has the disadvantage that variances can only be obtained for strata which contain an integral number of the basic strata—if the strata are all to be of the same shape the number must be a square. This disadvantage can be overcome by sub-stratifying the basic strata, with random selection of units from within these sub-strata. Thus in area sampling with square strata, if each basic stratum is subdivided into nine square sub-strata, with a minimum of two selected units per sub-stratum, square strata can be constructed with areas of $1, 1\frac{2}{9}, 2\frac{2}{9}, 4, 5\frac{4}{9}, 7\frac{1}{9}, 9, \dots$ times the area of the basic stratum. Separate analyses of variance will be required for the different sizes of strata, but these have certain elements in common.

When the variances have been calculated for certain sizes of strata an approximate variance-size relationship can be constructed by graphical means. It is often advantageous to plot the log-variance against the log-size. A straight line on this graph represents a variance law of the type $\sigma_z^2 = a z^b$, where z is the number of units per stratum and a and b are constants.

For the purpose of determining the size of sample required for a given accuracy it is better to plot $z \sigma_z^2$ against z . If N is the total number of units

in the population the number of units in a sample with two units per stratum will be $2N/z$, and we shall have $z\sigma_z^2 = 2NV(\bar{y})$. Thus the required size of the strata can be read off from the graph. With one unit per stratum the factor 2 is omitted.

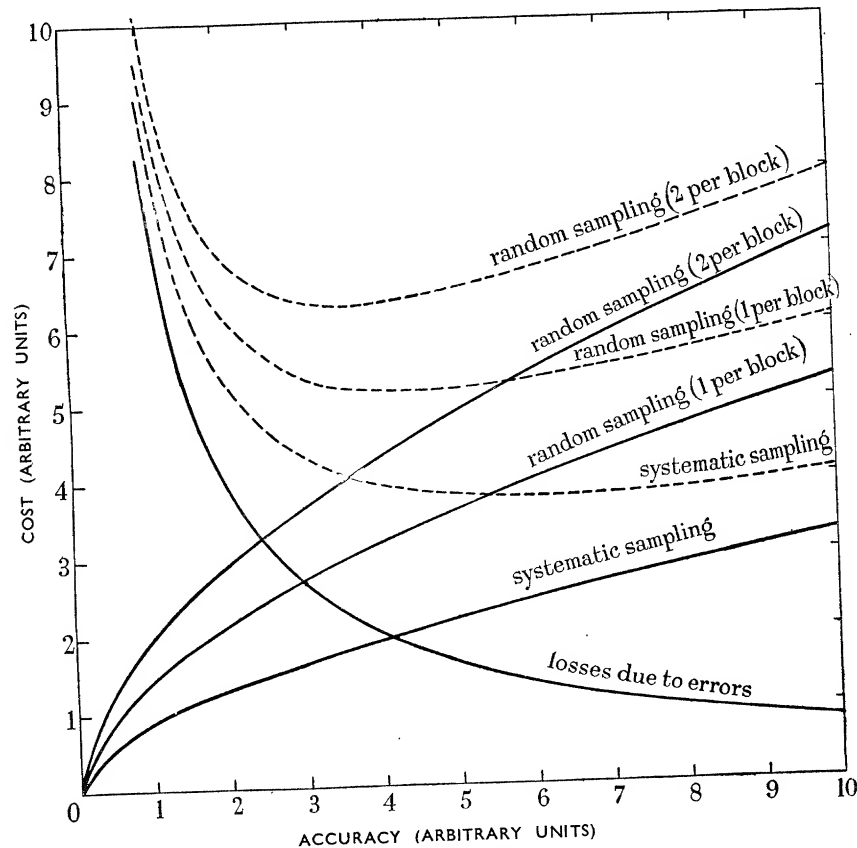


FIG. 8.15—RELATIONS BETWEEN COST AND ACCURACY IN SAMPLING FOR MEAN SOIL TEMPERATURE OVER A PERIOD, WITH STRATA OF VARYING SIZE

The mean temperature is estimated from temperatures taken (a) on two days selected at random from each stratum (block of days), (b) on one day selected at random from each stratum, (c) on days equally spaced throughout the period.

Reproduced by permission of the Royal Society (Yates, 1948, A).

It has been pointed out in Section 3.14 that systematic sampling, when used on the type of material for which it is suitable, is likely to have an error variance which is somewhat less than random sampling with one unit per stratum. In neither type of sampling can the sampling error be estimated with any certainty from the results of a single sample. In random sampling

with one unit per stratum, however, an objective estimate of error is possible if additional randomly located units are taken in certain of the strata, but in a systematic sample much more elaborate methods have to be used (Yates, 1948, A), and even then the estimates obtained are not fully objective.

It may be noted here that the common practice of estimating the error of a sample with one unit per stratum by combining the strata in pairs, will give an estimate of error which will generally be somewhat greater than the true error with strata of double the size and two units per stratum.

An example of the relation between the accuracy of sampling with two units per stratum, sampling with one unit per stratum, and systematic sampling, is given in Fig. 8.15. The curves (full lines) are based on the variances found for daily soil temperatures at 1 foot depth, each daily reading constituting a sampling unit. The cost scale is proportional to the number of units, and the accuracy scale gives the accuracy of the sample estimate of the mean soil temperature over a period. The curves themselves are based on relations for σ_2^2 of the type given above. The curve of losses due to errors and the broken curves will be referred to in Section 8.18.

The material is of the type in which the reduction in variance with reduction in size of strata may be expected to be considerable. This is brought out by the curves. The relative precisions of the three types of sampling are given by the intercepts of horizontal lines, which are in the ratio 1 : 1.75 : 4.24. The relative efficiencies are given by the reciprocals of the intercepts of vertical lines, which are in the ratio 1 : 1.36 : 2.22. This provides an illustration of the marked difference between relative precision and relative efficiency when reduction in the size of the strata results in a considerable reduction in variance per unit.

8.16 Efficiency in terms of cost

In the previous sections we have described how to determine the size of sample necessary to attain results of a given accuracy when various methods of sampling are used. We have also indicated how the relative efficiency (in terms of numbers of sampling units) of different sampling methods and variations in a given method may be judged. Minimization of the number of sampling units or amount of material included in the sample will not in general, however, give maximum efficiency in terms of cost. To attain this the sampling method must be so chosen that the total cost of the survey is minimized.

To minimize the total cost it is necessary to know the relative costs of the different operations. Exact evaluation of these costs is usually troublesome, and is only worth while if an extensive survey has to be undertaken, or if a series of surveys on similar material is contemplated. The matter is complicated by the fact that for many purposes it is the marginal cost of an additional unit, rather than the average cost per unit that is required. Nevertheless it is not difficult in the course of survey operations, or even in the course of a pilot survey, to obtain data which will serve to give rough estimates

of the main components of the costs. With the aid of such estimates the efficiency of further surveys of similar material can often be substantially improved.

When information on the costs of different types of operation is available it is possible to determine the values of the sampling fractions, etc., which for a given sampling method will give results of the required accuracy for the least cost. Such values may be termed the *optimal values*. It is also possible to determine which of two methods, each employed in the most efficient manner, will be the least costly.

The determination of optimal values of the sampling fractions, etc., requires minimization of the cost function, and will be dealt with in the next section. The choice between different methods when the optimal values of the sampling fractions, etc., are known, or when there are no variants of this type, can be obtained directly from the results of the previous sections.

Thus in the case in which there is the possibility of using supplementary information, if c_s represents the cost per unit of obtaining the supplementary information, c_o the marginal cost per unit when no supplementary information is obtained (these costs being taken to include the marginal costs of abstraction and computation), and C_1 represents the additional computational cost of utilizing the supplementary information (which apart from the above marginal cost per unit may be taken as broadly independent of the size of the sample), the total cost of a sample of n_s units with supplementary information, excluding elements of cost which are fixed for both methods, will be

$$C_s = C_1 + n_s(c_o + c_s)$$

and that for a sample of n_o units without supplementary information will be

$$C_o = n_o c_o$$

Under conditions in which the error variance is inversely proportional to the number of units in the sample, the two samples will be of equal accuracy when the numbers of units are in inverse ratio to the relative precision of the two methods with equal numbers. If the regression method of adjustment is used, therefore, and the sample is random,

$$n_s/n_o = 1 - \rho^2$$

where ρ is the true correlation coefficient between the main and the supplementary variates (estimate r).

Hence the use of supplementary information will be more efficient if

$$n_o c_o > C_1 + n_o(1 - \rho^2)(c_o + c_s)$$

i.e. if

$$n_o(c_o + c_s)\rho^2 > n_o c_s + C_1$$

If the cost of adjustment C_1 can be ignored this inequality becomes

$$c_s/c_o < \rho^2/(1 - \rho^2)$$

which is independent of n_o , and therefore of the accuracy required. Thus, for example, under these conditions, if $\rho = \frac{1}{2}$ the use of supplementary

information will be worth while if the cost of collection is less than one-third the cost of taking an equal number of additional units.

With $\rho = \frac{1}{2}$, however, the gain will not be marked unless the ratio of the costs is considerably less than $\frac{1}{3}$. With a ratio of $\frac{1}{6}$ the total costs will be in the ratio of 7 : 8 (minimum value, with zero value of the cost ratio, 3 : 4). With higher values of ρ the gains are more marked. With $\rho = \frac{3}{4}$ the two methods have equal cost when $c_s/c_o = 9/7$. In this case, when the ratio has the values $\frac{1}{2}$ and $\frac{1}{4}$ the ratio of the total costs will be 21 : 32 and 35 : 64 respectively (minimum value 7 : 16).

8.17 Minimization of the cost function

When the sampling fractions, etc., of a method of sampling are not fully determined by the accuracy required in the results, the optimal values can be determined by minimizing the cost function.

The total cost can usually be expressed as a linear function of the numbers of sampling units $n_1, n_2, \dots$ in the various strata, etc., at least to a first approximation, using marginal costs. The simplest procedure is then to add a multiple K of this linear function to the expression in terms of $n_1, n_2, \dots$ for the variance of the required estimate, and differentiate the resultant expression with respect to $n_1, n_2, \dots$ in turn. This minimizes the variance for fixed cost, which is equivalent to minimizing the cost for fixed variance. The exact procedure will be apparent from the first of the cases treated below.

(a) Variable sampling fraction

If the marginal cost of taking an additional unit of the i th stratum is c_i , the total cost C , omitting constant elements, is given by

$$C = \sum c_i n_i$$

Hence

$$\begin{aligned} V(Y) &= \sum \sigma_i^2 (1 - f_i) N_i^2 / n_i \\ &= \sum \sigma_i^2 (1/n_i - 1/N_i) N_i^2 + K (\sum c_i n_i - C) \end{aligned} \quad (8.17.a)$$

Differentiating with respect to the n_i and equating to zero, we have the t equations ($i = 1, 2, \dots, t$)

$$-\sigma_i^2 N_i^2 / n_i^2 + K c_i = 0$$

Hence, since $n_i/N_i = f_i$,

$$\frac{f_1}{\sigma_1/\sqrt{c_1}} = \frac{f_2}{\sigma_2/\sqrt{c_2}} = \dots = \frac{1}{\sqrt{K}} \quad (8.17.b)$$

Thus the optimal sampling fractions are proportional to $\sigma_i/\sqrt{c_i}$. This is an extension of the formula already given in Section 3.5.

The actual values of the sampling fractions required to attain a given accuracy can be obtained by substituting for the f_i in equation 8.17.a and solving for K . This gives

$$(\sqrt{K}) \sum N_i \sigma_i \sqrt{c_i} = V(Y) + \sum N_i \sigma_i^2$$

If any of the f_i are equal to unity the corresponding terms must be omitted from both sides of the equation.

(b) *Two-phase sampling*

If c_1 represents the cost per unit of obtaining the first-phase information, c_2 the additional cost per unit of obtaining the second-phase information, and there are n_1 first-phase units, of which n_2 are included in the second phase, the total cost, apart from constant elements, will be given by

$$C = n_1 c_1 + n_2 c_2$$

When the methods of sampling and estimation are such that the effective variances of the estimates at each phase (apart from corrections for finite sampling) are inversely proportional to the numbers of units, from the results of Section 8.7 we have

$$V(\bar{y}) = \frac{1-f_1}{n_1} \sigma_1^2 + \frac{1}{n_2} \left(1 - \frac{n_2}{n_1}\right) \sigma_2^2 \quad (8.17.c)$$

Following the above procedure, we find

$$\frac{n_2^2}{n_1^2} = \frac{c_1}{c_2} \cdot \frac{\sigma_2^2}{\sigma_1^2 - \sigma_2^2} = \frac{c_1}{c_2} \cdot \frac{\kappa^2}{1 - \kappa^2} \quad (8.17.d)$$

where $\kappa = \sigma_2/\sigma_1$. The values of n_1 and n_2 required for a given accuracy can be obtained by substituting for n_2 in terms of n_1 in $V(\bar{y})$.

(c) *Two-phase point sampling*

We will only consider the special case arising in crop estimation (Examples 6.16.b and 7.15).

If the acreages are determined from n_0' points and the yields per acre are determined on fields of the crop in question in which n of these points fall, and if c' is the cost of visiting the field to determine the nature of the crop, and c the additional cost of a yield determination, we have, when a proportion p of the area is under the crop and the mean yield per acre is $\bar{r}$,

$$C = c' n_0' + c n$$

$$V(Y)/Y^2 = q/p n_0' + V(r)/n \bar{r}^2 \quad (8.17.e)$$

Hence

$$\frac{n^2}{n_0'^2} = \frac{p V(r) c'}{q \bar{r}^2 c} \quad (8.17.f)$$

The values of n_0' and n are best obtained by substitution for n in terms of n_0' in the equation for $V(Y)$.

(d) *Two-stage sampling*

If c' is the cost per first-stage unit, and c'' the additional cost per second-stage unit, the total cost is given by

$$C = n' c' + n c''$$

With a random or stratified random sample with uniform sampling fraction and equal numbers of second-stage units per first-stage unit, $V(\bar{y})$ is given by formula 8.11.c. Thus

$$V(\bar{y}) = U/n' + V/n + \text{const.},$$

where

$$U = \sigma_0'^2 - \sigma''^2/N'', \quad \text{and} \quad V = \sigma''^2.$$

Following the previous procedure, we find

$$n^2/n'^2 = n''^2 = V c' / U c'' \quad (8.17.g)$$

In other words the number of second-stage units per first-stage unit is independent of the accuracy required. The values of n' and n required for any given accuracy can be obtained by substitution in the equation for $V(\bar{y})$.

The same formulæ hold for any form of two-stage sampling in which $V(\bar{y})$ can be written in the above form. Thus stratification with a variable sampling fraction at the second stage is covered, provided none of the sampling fractions are unity.

(e) *Two-stage sampling with probability proportional to size at the first stage*

The solution of the case of sampling from within strata with probability proportional to size of unit at the first stage follows similar lines. We find that if the cost per second-stage unit is the same for all first-stage units in all strata, one condition for minimum cost is that the second-stage sampling fractions are so chosen that the overall sampling fraction is uniform. Thus the use of a uniform overall sampling fraction, which is computationally convenient, is justified on grounds of minimum cost. The assumption of constant cost per second-stage unit will not in fact hold for the component of cost due to travel, since the same number of second-stage units will be taken from any selected unit of a stratum, and consequently the travel cost per unit will be greater for the larger units. This, however, is not likely to reduce the efficiency greatly unless travel costs at the second stage are very large.

The relation between the first-stage and second-stage sampling can also be very simply expressed. In the case considered in Section 8.13, in which the size of the first-stage units is represented by the number of second-stage units, if all the second-stage variances are equal we may put $\sigma_{ri0}'^2 - \sigma''^2 N_i'/N_i = U_i$ and $\sigma''^2 = V$, the costs per first and second-stage unit being taken as c_i' and c'' . We then have

$$f = \frac{V N + \sqrt{(V/c'')} \sum N_i \sqrt{(U_i c_i')}}{V(Y) + \sum N_i^2 U_i / N_i'}$$

$$n_i'^2 = f^2 N_i^2 U_i c'' / V c_i'$$

Since the number of second-stage units n_i'' per first-stage unit in stratum i is the same for all selected units, and independent of which particular units are selected, the above equations give $n_i''^2 = V c_i' / U_i c''$, as before.

(f) *Two-stage sampling of farms and fields*

Any fully general treatment is difficult owing to the fact that the numbers of fields per farm carrying a given crop are usually small, and consequently what would otherwise be the optimal values of the second-stage sampling fractions will give numbers of fields per farm which are not only non-integral, but which will in many cases be less than unity.

If the numbers of fields per farm are sufficiently large for this source of disturbance to be neglected, the optimal values of the sampling fractions can be simply expressed.

In order to standardize the notation we may replace the between-farms component of variance U_2 , as defined in Section 8.12, by U' , and the between-fields within-farms component U_1' by U'' . The cost of visiting a farm may be taken as c' , and that of sampling a field as c'' .

We will consider the case in which the farms are divided into size-groups with fields which have mean areas $\bar{a}_1, \bar{a}_2, \dots$. We will further assume that the mean acreages per field within a size-group of farms of 1, 2, 3, . . . fields are the same, and that $V(a_i)/\bar{a}_i^2$ is constant for all size-groups and for all numbers of fields within a size-group.

In the first place we find that in this case the second-stage sampling fractions within a size-group should all have the same value, which is given by

$$f_i'^2 = \frac{U''}{U'} \frac{c'}{c''} \frac{N_i'}{[N_i']_2} \quad (8.17.h)$$

where $N_{i1}', N_{i2}', N_{i3}', \dots$ are the numbers of farms in the group with 1, 2, 3, . . . fields respectively, and $[N_i']_2 = N_{i1}' + 4 N_{i2}' + 9 N_{i3}' + \dots$. The ratios of the first-stage sampling fractions are given by

$$\frac{f_1'^2}{\bar{a}_1^2 [N_1']_2 / N_1'} = \frac{f_2'^2}{\bar{a}_2^2 [N_2']_2 / N_2'} = \dots = k^2 \text{ say} \quad (8.17.i)$$

These equations will serve to give first approximations to the relative sampling fractions. The relative efficiency of different variants which are practically applicable can then be tested by use of the expression for the variance of the weighted mean given in Section 8.12. It will usually be sufficient to use the mean acreage for each size-group in evaluating the weights, but in evaluating the actual size of sample required the factors $\bar{a}_i^2$ should be replaced by $\bar{a}_i^2 + V(a_i)$.

Example 8.17.a

Determine, from the data of Examples 6.12.b and 7.12.b, the optimal proportion of sample plots to eye estimates on conifer stands in a two-phase sampling scheme in which eye estimates only are made at the first phase, when the cost of visiting a stand and making an eye estimate is $\frac{1}{10}$ the additional cost of measuring a sample plot.

Consideration of this problem would be relevant if it were possible to demarcate and classify the stands into conifers of over 20 years of age, etc., from aerial photographs. In this case, if the sampling of stands is with probability proportional to size, the variances of the volumes per acre, given in Example 7.12.b, will be required. We then have, with the regression method of estimation,

$$\begin{aligned} s_1^2 &= 4803 & s_2^2 &= 3579 \\ \kappa^2 &= 0.745 & \kappa^2/(1 - \kappa^2) &= 2.92 & c_1/c_2 &= 1/10 \\ n_2/n_1 &= \sqrt{(2.92/10)} & &= 0.540 \end{aligned}$$

Thus sample plots should be taken on about one-half the stands which are visited.

There is, however, in this case no appreciable gain by the use of two-phase sampling. If n_2' is the number of sample plots required if no eye estimates are made we have, for equal variance,

$$\frac{1}{n_2'} \sigma_1^2 = \frac{1}{n_1} \sigma_1^2 + \frac{1}{n_2} \left(1 - \frac{n_2}{n_1}\right) \sigma_2^2$$

which gives

$$\begin{aligned} n_2/n_2' &= n_2/n_1 + (1 - n_2/n_1) \kappa^2 \\ &= 0.540 + 0.460 \times 0.745 = 0.883 \\ n_1/n_2' &= 1.64 \end{aligned}$$

Thus with two-phase sampling

$$C/n_2' c_1 = 1.64 + 10 \times 0.883 = 10.47$$

and with single-phase sampling $C/n_2' c_1$ lies between 10 and 11, depending on the saving due to the omission of the eye estimates on the stands that are visited.

One of the reasons why the use of eye estimates is here of little value is that the determination of the volumes of individual stands by means of a single sample plot per stand is very inaccurate. If more sample plots per stand were taken the overall variance of y would be reduced, while the covariance of y and x and the variance of x would remain unaltered. Under these circumstances two-phase sampling would be more advantageous. Given information on the within-stand variance of the sample plots and the cost of taking different numbers of sample plots from a stand, the optimal number of sample plots per stand could be determined.

Example 8.17.b

If in the crop survey of Examples 6.16.b and 7.15 the additional cost of crop-cutting in order to obtain an estimate of yield is 20 times the cost of visiting a sampling point to ascertain the nature of the crop, calculate the optimal ratio of the number of yield determinations to total sample points, and the

number of points required to give an estimate of the total yield with a standard error of 5 per cent.

From the results already given we have

$$\begin{aligned} p &= 2,202/33,255 = 0.0662 & q &= 0.9338 \\ \bar{r} &= 15.7 & V(r) &= 3.5^2 & V(r)/\bar{r}^2 &= 0.0497 \end{aligned}$$

Hence, from equation 8.17.f,

$$\frac{n}{n_0'} = \sqrt{\frac{0.0662 \times 0.0497}{0.9338 \times 20}} = 0.0133$$

Substituting in equation 8.17.e,

$$\begin{aligned} (.05)^2 n_0' &= .9338/0.0662 + .0497/0.0133 \\ n_0' &= 7140 & n &= 95 & n' &= 473 \end{aligned}$$

The large number of points that have to be visited to ascertain the crop is accounted for by the small fraction of the total land area under crop. If the crop is an important one it will occupy a considerably larger fraction of the cultivated area, and if, therefore, the non-cultivated areas can be excluded, the total number of points required will be considerably reduced. Alternatively sparsely cultivated areas may be sampled with a lower intensity.

It is also worth noting that if several crops have to be surveyed it will probably be possible to make the acreage determinations of all crops simultaneously prior to the crop-cutting work. This will alter the above cost relationships. The general case can be dealt with by minimization of the combined cost function. In the simple case in which there are a number of crops each occupying the same area and having the same variance and cost relationships, and in which the same accuracy is required for each crop, the above solution holds, the cost of the acreage determinations being spread equally over all the crops, and adjustment of c being made for the cost of revisits. Thus in the above example, with 5 crops and a cost of revisit per point of double the original cost (owing to wider dispersion), all that is necessary is to put $c/c' = 110$. We then find $n_0' = 9160$, $n = 52$, $n' = 606$.

Example 8.17.c

If county lists of farms are not available, and if the cost of the construction of a list of the farms of a parish is 10 times the cost of visiting a single farm within the parish and ascertaining the wheat acreage, determine the optimal sampling fractions at the first and second stage which will give estimates of the acreage of wheat having a standard error of ± 4500 (*i.e.* approximately 10 per cent.), using the methods of sampling followed in samples B_1 and B_2 of Hertfordshire farms.

From the equation for $V(Y)$ in terms of n' and n given in Example 8.11.a and equation 8.17.g, we have

$$n'' = \sqrt{(2524 \cdot 2 \times 10/80 \cdot 71)} = 17 \cdot 7$$

Hence, for the required accuracy,

$$4 \cdot 5^2 = 80 \cdot 71/n' + 2524 \cdot 2/17 \cdot 7 n' - 1 \cdot 8982$$

$$n' = 10 \cdot 1$$

$$n = 178 \cdot 8$$

Similarly, from Example 8.11.b,

$$n'' = 7 \cdot 0 \quad n' = 11 \cdot 2 \quad n = 78 \cdot 4$$

For the reasons already given these latter values are approximate. When proper allowance is made for the changes in sampling fractions at the second stage a somewhat smaller sample will be found to be necessary.

Example 8.17.d

From the data of Table 6.19.a determine suitable sampling fractions for a crop-estimation scheme for sugar beet, on the assumption that variances between fields on the same farm, and between farms, are the same as those found for wheat in Section 8.12, and that the cost of visiting a farm is (a) equal to, and (b) twice that of sampling a field.

From Table 6.19.a, including farms not growing sugar beet on old arable land, we obtain the following values:

	n_i'	$[n_i']_2$	$\bar{a}_i$
Small farms	30	51	3.1
Medium farms	47	224	8.4
Large farms	14	121	12.3

Taking the estimates of $N_i'/[N_i']_2$ given by $n_i'/[n_i']_2$, formulæ 8.17.h and 8.17.i give the following values for f_i'' and f_i' :

	f_i''	f_i''	f_i'/k
Small farms	$c_1 = c_2$	$c_1 = \frac{1}{2} c_2$	All c
Medium farms	0.86	1.21	4.0
Large farms	0.51	0.72	18.3
	0.38	0.54	36.2

Thus when $c_1 = c_2$ we may consider variants on the following scheme. For small farms sample all fields; for medium farms sample one field from farms growing 1 or 2 fields, two fields from farms growing 3 or 4 fields, etc.; for large farms sample one field from farms growing 1, 2, 3 or 4 fields, two fields from farms growing 5, 6 or 7 fields, etc.; sample farms in the proportions 1:4:8.

A similar scheme can be drawn up for the case when $c_1 = \frac{1}{2} c_2$. In this case, since f_1'' is greater than unity, some increase in the proportion of small farms included may be advisable.
Further investigation is left to the reader.

8.18 Losses due to errors

We have so far considered the minimization of the cost of a survey when a given accuracy is required in the results. Since, however, errors in the results themselves give rise to losses when these results are used as a basis for further action, the accuracy should itself be determined in such a manner that the sum of the cost of the survey and the expected losses due to the resultant errors is minimized.

If the loss due to an error Z in an estimate Y is equal to $a Z^2$, where a is a constant, the average loss in a series of samples of the same size and type in which the estimates are free from bias will be $a V(Y)$, whatever the actual form of the distribution function of the errors.

When the loss due to an error is proportional to the square of the error, therefore, minimization of the sum of the cost C of a survey and the average loss due to errors requires minimization of the function

$$C + a V(Y)$$

the sampling method and size of sample being so chosen that $V(Y)$ has its minimum possible value for the cost C .

Under these circumstances $V(Y)$ will always be expressible as a function of C . In many cases, as we have seen, this function is of the form

$$V(Y) = \frac{h}{C - C_0} - k$$

where h and k are constants depending on the population which is being surveyed, and C_0 is the overhead or constant component of cost which is independent of the size of the survey.

In this case the minimum value will be attained when

$$(C - C_0)^2 = ah$$

We then have

$$V(Y) = \frac{\sqrt{h}}{\sqrt{a}} - k$$

This implies that the more accurate the results that can be obtained with a given cost, *i.e.* the smaller the value of h , the higher should be the accuracy aimed at with a given loss function. The value of the cost-plus-loss function at the minimum is in fact $2C - C_0 - ak$. Any saving due to increased accuracy should therefore be divided equally between reduction in the cost of the survey and reduction in the loss due to errors. Equally if the loss due to a given error is multiplied by a factor λ , the funds devoted to the survey (excluding overhead) should be multiplied by $\sqrt{\lambda}$.

The same general conclusions hold when the variance-cost relation is of a more complicated form than that given above. A case of this type is illustrated in Fig. 8.15, in which a loss curve of the form $aV(Y)$ has been inserted. We see that with the more accurate methods of sampling the minima of the cost-plus-loss functions (shown by broken lines) are attained when both the cost of the sampling and the loss due to errors are less than with the less accurate methods.

Other loss functions will lead to more complicated expressions for the average loss. The most general loss function which is capable of relatively simple expression in terms of $V(Y)$ is that in which the loss due to a positive error is equal to aZ^b , and that due to a negative error is $a'(-Z)^b$, a , a' and b being constants. Provided the distribution of errors has the same form for all values of $V(Y)$, the average loss is then equal to $a''\sigma_0^b$, where $\sigma_0^2 = V(Y)$ and a'' has a value which is a linear function of a and a' . The actual linear function can only be determined if the distribution function of the errors is known. In general terms, if the distribution function of the errors of Y is $f_1(z) dz$, where $z = Z/\sigma_0$, we have

$$a'' = a \int_0^{+\infty} z^b f_1(z) dz + a' \int_{-\infty}^0 (-z)^b f_1(z) dz$$

If the distribution of the errors is normal, $f_1(z) dz$ will be of the form given in Section 7.3, with $\sigma = 1$. The two integrals will in this case (as in any symmetrical distribution) be equal. Their values for any value of b can be obtained from existing tables*, those for $b = 1.0, 1.25, 1.5, 1.75, 2.0$ being 0.3989, 0.4097, 0.4300, 0.4599, 0.5 respectively. It must be emphasized, however, that the distribution of sampling errors is frequently not sufficiently normal for the use of these values to be justified. In such cases, also, the form of the distribution may be expected to change with change in the size of the sample.

With this more general loss function we require to minimize the function

$$C + a'' \left(\frac{h}{C - C_0} - k \right)^{b/2}$$

which will be minimum when

$$(C - C_0)^2 \left(\frac{h}{C - C_0} - k \right)^{1-b/2} = \frac{1}{2} a'' b h$$

This equation can easily be solved by trial and error, or directly if k can be neglected, as will be the case when all sampling fractions are small. The same general conclusion that the accuracy should be increased with a more accurate method of survey still holds.

When $V(Y)$ is a more complicated function of C (possibly only determined in numerical form) the minimum of the cost-plus-loss function can itself be determined by trial and error.

* Tables of the Gamma function.

8.19 Concluding remarks

The preceding sections give an indication of the ways in which the efficiencies of different sampling methods can be compared, and the techniques of determining the optimal sampling fractions, size of sample required for a given accuracy, etc. It has been further shown that the accuracy which should be aimed at is itself related to the losses resulting from errors in the survey results.

Since determination of the optimal accuracy from the expected losses due to errors demands knowledge of the loss-function, it will chiefly be of relevance when action in the economic sphere has to be based on the results of the survey. An error in the estimate of the yield of a crop, for instance, may require changes in an import programme, or may lead to wastage, and the resultant additional costs may be assessable, at least roughly. The losses due to errors in estimates provided by surveys of the research and investigational type can scarcely be assessed. Indeed, it is usually impossible to give any quantitative estimate in monetary terms of the value of the information provided by such surveys. The decision to undertake the survey, and the accuracy aimed at, must then be a matter of judgment on the part of those who require the information, and those who are concerned with the allocation of resources.

Even if the optimal accuracy cannot be quantitatively determined, arbitrary decisions on the accuracy required should as far as possible be avoided. Before any decision as to accuracy is taken, estimates should be prepared of the costs of obtaining results of differing degrees of accuracy, and these estimates should be considered in relation to the purposes for which the results are required.

Minimization of costs can of course be carried out whether or not a loss-function is available. In this chapter we have only considered this minimization when a single quantity requires estimation. In most censuses and surveys such treatment would be an over-simplification. A number of quantities will require to be estimated, frequently for many domains of study. It may then be necessary to carry out a more elaborate investigation, minimizing the cost for defined accuracies of all the estimated quantities. Alternatively, if loss-functions are available for all of these quantities, the combined cost-plus-loss function can be minimized. Frequently, however, one of the quantities is of dominant importance, and the situation is such that when adequate accuracy is attained on this quantity the remaining quantities are determined with more than the required accuracy. In this case minimization can be conducted solely with reference to this quantity.

Many of the examples worked out in this chapter are based on very small amounts of data, and the conclusions reached on the relative efficiencies of different methods, even in the particular circumstances of the chosen examples, must therefore be treated with reserve. These examples are, in fact, merely intended to illustrate the computational procedures, and bring out the various points that have to be taken into account when making calculations of relative efficiencies, optimal sampling fractions and size of sample. They are in no

way intended as general investigations into the relative efficiency of the different methods.

On the other hand, it should be borne in mind that no very exact determinations of the optimal sampling fractions and size of sample are required in the practical planning of surveys. If the values adopted are somewhere near the optimal the total cost, or the total of costs-plus-losses, will be very near the minimum.

We must also not be deterred from undertaking a survey by the fact that there is little information on which to base exact planning of the sampling methods. As we have seen, surveys themselves provide information which will enable future surveys on similar material to be more efficiently planned. In surveys on relatively unknown material one of the points to be kept in mind in the planning is that information will be required both on variances and on costs. Equally, if preliminary rough estimates are required, pilot investigations can be designed so as to provide such estimates, as well as information on which to base the planning of a larger survey.

The study of the relative efficiency of different sampling methods depends not so much on having a large amount of data as on having data which are relevant to the methods concerned. Thus the small pilot investigation on the estimation of wheat yields by sampling methods described in Section 8.12 was of sufficient size to give estimates of both the field-to-field and farm-to-farm components of variance with all necessary accuracy. On the other hand, in the Survey of Fertilizer Practice—although a very large amount of data has now been accumulated—it is impossible to determine the field-to-field components of variation of the fertilizer dressings, since only one old- and one new-arable field of each crop was taken on each farm. This must be regarded as a defect in the planning of this survey, which could have been remedied had a pair of fields been taken for the various crops on a small proportion of the farms. Incidentally, lack of this information has also prevented any consideration of the question of the extent to which individual farmers vary their fertilizer practice from field to field of the same crop.

Given the necessary data, the increase in the efficiency of survey methods requires proper statistical investigations of the types outlined in this chapter. The need for thorough investigation of the efficiencies of different sampling methods in different circumstances is great, and it is to be hoped that many more will be made and reported in the future. Such investigations are often neglected because, once a survey has been completed, the question of whether it could have been carried out more efficiently is largely historical as far as that survey is concerned. One of the reasons why both the theory and practice of sample censuses and surveys has made rapid advances in recent years is that permanent organizations—often part of, or attached to, research statistical institutes—have been set up in a number of countries. These organizations have been actively engaged both in the planning and the execution of surveys covering various fields of enquiry. Consequently they have not only had access to the necessary data, or the means of collecting it, but they have also

had a continuing interest in investigations of efficiency, and a body of workers who have both the training and experience to carry out these investigations.

Further progress may be expected on the same lines. In particular the problems that arise in censuses and surveys of undeveloped areas will be likely to receive very much more thorough investigation when more centres which are actively concerned with the planning and execution of surveys in these areas are developed. Only in this way will a body of experience be built up which is relevant to the special problems of such surveys.

TABLE A1—RANDOM NUMBERS

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27
18 18 07 92 46	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 42 06 87 98
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04
56 62 18 37 35	96 83 50 87 75	97 12 25 93 47	70 33 24 03 54
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27
16 90 82 66 59	83 62 64 11 12	67 19 00 71 74	60 47 21 29 68
11 27 94 75 06	06 09 19 74 66	02 94 37 34 02	76 70 90 30 86
35 24 10 16 20	33 32 51 26 38	79 78 45 04 91	16 92 53 56 16
38 23 16 86 38	42 38 97 01 50	87 75 66 81 41	40 01 74 91 62
31 96 25 91 47	96 44 33 49 13	34 86 82 53 91	00 52 43 48 85
66 67 40 67 14	64 05 71 95 86	11 05 65 09 68	76 83 20 37 90
14 90 84 45 11	75 73 88 05 90	52 27 41 14 86	22 98 12 22 08
68 05 51 18 00	33 96 02 75 19	07 60 62 93 55	59 33 82 43 90
20 46 78 73 90	97 51 40 14 02	04 02 33 31 08	39 54 16 49 36
64 19 58 97 79	15 06 15 93 20	01 90 10 75 06	40 78 78 89 62
05 26 93 70 60	22 35 85 15 13	92 03 51 59 77	59 56 78 06 83
07 97 10 88 23	09 98 42 99 64	61 71 62 99 15	06 51 29 16 93
68 71 86 85 85	54 87 66 47 54	73 32 08 11 12	44 95 92 63 16
26 99 61 65 53	58 37 78 80 70	42 10 50 67 42	32 17 55 85 74
14 65 52 68 75	87 59 36 22 41	26 78 63 06 55	13 08 27 01 50

This table forms part of a larger table of random numbers given in *Statistical Tables for Biological, Agricultural and Medical Research* by R. A. Fisher and F. Yates, Oliver & Boyd, Edinburgh (3rd edition, 1948), and is reproduced by kind permission of the senior author and the publishers.

TABLE A2—THE NORMAL DISTRIBUTION

Probability of obtaining deviations (positive or negative) greater
than given multiples of the standard deviation

Deviation z/σ	Probability P	Deviation z/σ	Probability P	Deviation z/σ	Probability P
0.0	1.0000	1.0	.3173	2.0	.0455
0.1	.9203	1.1	.2713	2.1	.0357
0.2	.8415	1.2	.2301	2.2	.0278
0.3	.7642	1.3	.1936	2.3	.0214
0.4	.6892	1.4	.1615	2.4	.0164
0.5	.6171	1.5	.1336	2.5	.0124
0.6	.5485	1.6	.1096	2.6	.0093
0.7	.4839	1.7	.0891	2.7	.0069
0.8	.4237	1.8	.0719	2.8	.0051
0.9	.3681	1.9	.0574	2.9	.0037
1.0	.3173	2.0	.0455	3.0	.0027

BIBLIOGRAPHY ON SAMPLING

This bibliography has been drawn up by Mr. D. R. Read. It is based on a bibliography prepared by the Statistical Office of the United Nations, though not all the references in the United Nations' bibliography are included here.

The papers have been classified under the following heads :—

- (A) Theory and methods.
- (B) Machine methods.
- (C) Population censuses.
- (D) Sociology, nutrition, health, etc.
- (E) Opinion surveys and market research.
- (F) Economics : surveys of industry, censuses of production, labour force, etc.
- (G) Agricultural economics and farm practice.
- (H) Crop estimation and forecasting, etc.
- (I) Forestry and land utilization surveys.
- (J) Estimation of wild populations.

Since a single paper does not necessarily deal with only one subject, the subject classification must be taken as approximate only. A certain amount of general theory, for example, will be found in papers primarily dealing with special applications. In some instances where the original paper could not be consulted the classification has been made from the title and journal. Papers by the same author may be found under more than one heading, but papers by more than one author are indexed in the section concerned under the name of each author, so as to avoid difficulty in tracing all papers by a given author.

BOOKS

General

- BAEHNE, G. W. (1935). "Practical applications of the punched card method in colleges and universities." New York : Columbia University Press.
- BLANKENSHIP, A. (1943). "How to conduct consumer, and opinion research" (2nd. edn., 1945). New York : Harpers.
- CANTRIL, H. (1944). "Gauging public opinion." Princeton University Press.
- CHURCHMAN, C. W., ACKOFF, R. L., and WAX, M. (1947). "Measurement of consumer interest." Philadelphia : University of Philadelphia Press.
- FISHER, R. A. (1925). "Statistical methods for research workers" (10th edn., 1946). Edinburgh : Oliver & Boyd.
- (1935). "The design of experiments" (4th edn., 1947). Edinburgh : Oliver & Boyd.
- HARTKEMEIER, H. P. (1942). "Principles of punch-card machine operation." New York : Thomas Y. Crowell.
- KENDALL, M. G. (1943, 1946). "The advanced theory of statistics." Vol. I (3rd edn., 1947) & Vol. II (2nd edn., 1948). London : Griffin.
- PEATMAN, J. G. (1947). "Descriptive and sampling statistics." New York : Harpers.
- RHODES, E. C. (1933). "Elementary statistical methods" (8th edn., 1948). London : Routledge.

- SCHUMACHER, F. X., and CHAPMAN, R. A. (1942). "Sampling methods in forestry and range management." Duke University, Durham, North Carolina.
- SNEDECOR, G. (1937). "Statistical methods" (4th edn., 1946). The Collegiate Press, Ames, Iowa.
- THIONET, P. (1946). "Méthodes statistiques modernes des Administrations Fédérales aux Etats-Unis." Paris: Herman et Cie.
- YULE, G. U. (1911), and KENDALL, M. G. "An introduction to the theory of statistics" (13th edn., 1945). London: Griffin.

Reports on Surveys, etc.

- BOWLEY, A. L. (1930-1935). "New survey of London life and labour." Vols. I-IX. London: P. S. King.
- JONES, D. CARADOG (1934). "The social survey of Merseyside." London: Hodder & Stoughton.
- ROWNTREE, B. SEEBOHM (1901). "Poverty: a study of town life" (4th edn., 1908). London: Macmillan.
- TOUR, H. (1939). "The standard of living in Bristol: A preliminary report of the work of the University of Bristol Social Survey." Bristol: Arrowsmith.
- U.S. BUREAU OF CENSUS (1947). "A chapter in population sampling." U.S. Govt. Printing Office, Washington, D.C.

Tables

- FISHER, R. A., and YATES, F. (1938). "Statistical tables for biological, agricultural and medical research" (3rd edn., 1948). Edinburgh: Oliver & Boyd.

PAPERS

A. THEORY AND METHODS

- ALTSCHUL, H. (1913). "Studie über die Methode der Stichprobenerhebung." *Arch. Rass.- u. Ges. Biol.*, **10**, 110-152.
- ANDERSON, P. H. (1942). "Distributions in stratified sampling." *Ann. Math. Statist.*, **13**, 42-52.
- ARMITAGE, P. (1947). "A comparison of stratified with unrestricted random sampling from a finite population." *Biometrika*, **34**, 273-280.
- BERNERT, E. H. (1945). See HAGOOD, M. J.
- BOSE, C. (1943). "Note on the sampling error in the method of double sampling." *Sankhya*, **6**, 329-330.
- BOWLEY, A. L. (1926). "Measurement of the precision attained in sampling." *Bull. Inst. Int. Statist.*, **22**, (1), 1-62.
- COCHRAN, W. G. (1939). "The use of analysis of variance in enumeration by sampling." *J. Amer. Statist. Ass.*, **34**, 492-510.
- (1942). "Sampling theory when the sampling units are of unequal sizes." *J. Amer. Statist. Ass.*, **37**, 199-212.
- (1946). "Relative accuracy of systematic and stratified random samples for a certain class of populations." *Ann. Math. Statist.*, **17**, 164-177.
- (1948). "Recent developments in sampling theory in the United States" (Abstract). *Econometrica*, **16**, 71-72.
- CORNELL, F. G. (1947). "A stratified random sample of a small finite population." *J. Amer. Statist. Ass.*, **42**, 523-532.
- CORNFIELD, J. (1942). "On certain biases in samples of human population." *J. Amer. Statist. Ass.*, **37**, 63-68.
- (1944). "On samples from finite populations." *J. Amer. Statist. Ass.*, **39**, 236-239.
- COX, G. M. (1935). See SNEDECOR, G. W.
- CRAIG, A. T. (1939). "On the mathematics of the representative method of sampling." *Ann. Math. Statist.*, **10**, 26-34.

- DEMING, W. E. and STEPHAN, F. F. (1940). "On a least squares adjustment of a sampled frequency table when the expected marginal totals are known." *Ann. Math. Statist.*, **11**, 427-444.
- (1945). "On training in sampling." *J. Amer. Statist. Ass.*, **40**, 307-316.
- FRANKEL, L. R. (1939). See STOCK, J. S.
- GALVANI, L. (1929). See GINI, C.
- GANGULI, M. (1941). "A note on nested sampling." *Sankhya*, **5**, 449-452.
- GHOSH, B. (1947). "Bias introduced by changing the system of stratification." *Calcutta Statist. Ass. Bull.*, No. 1, 43-45.
- (1947). "Double sampling with many auxiliary variates." *Calcutta Statist. Ass. Bull.*, No. 2, 91-93.
- GINI, C. and GALVANI, L. (1929). "Di una applicazione del metodo rappresentativo all'ultimo censimento italiano della popolazione" (1° dicembre 1921). *Annali di Statistica*, VI, **4**, 1-107.
- GRAVELL, W. (1923). "Die Not der Statistik und die representative Methode." *Allg. Statist. Archiv*, **13**, 345-346.
- (?) . "Die representative Methode." *Deutsch. Statist. Zent.*, **15**, 5.
- HAGOOD, M. J., and BERNERT, E. H. (1945). "Component indexes as a basis for stratification." *J. Amer. Statist. Ass.*, **40**, 330-341.
- HANSEN, M. H. and HURWITZ, W. N. (1943). "On the theory of sampling from finite populations." *Ann. Math. Statist.*, **14**, 333-362.
- HENDRICKS, W. A. (1948). "Mathematics of sampling." Virginia Agric. Exp. Sta., Special Bulletin.
- HURWITZ, W. N. (1943). See HANSEN, M. H.
- JENSEN, A. (1923). "Arbejdsbesparende Metoder i Statistikken." *Nordisk Statist. Tidsskrift*, Stockholm, **2**, 409-434.
- (1926). "The application of the representative method." *Bull. Inst. Int. Statist.*, **22**, 58-60.
- (1926). "Report on the representative method in statistics." *Bull. Inst. Int. Statist.*, **22**, 359-380.
- (1926). "The representative method in practice." *Bull. Inst. Int. Statist.*, **22**, 381-439.
- (1928). "Purposive selection." *J. R. Statist. Soc.*, **91**, 541-547.
- KIAER, A. N. (1915). "Sur les méthodes représentatives ou typologiques appliquées à la Statistique." *Bull. Inst. Int. Statist.*, **11**, 180-214.
- (1917). "Sur les méthodes représentatives ou typologiques appliquées à la Statistique." *Bull. Inst. Int. Statist.*, **13**, 66-101.
- LUCHT, J. (1922). "Die representative Methode in der Statistik." *Zeits. preuss. statist. Landesamts*, **62**, 122-141.
- MADON, W. & L. (1944). "On the theory of systematic sampling." *Ann. Math. Statist.*, **15**, 1-24.
- (1946). "Systematic sampling and its relation to other sampling designs." *J. Amer. Statist. Ass.*, **41**, 204-217.
- MAHALANOBIS, P. C. (1944). "On large scale sample surveys." *Roy. Soc. Phil. Trans.*, B, **231**, 329-451.
- (1946). "Recent experiments in statistical sampling in the Indian Statistical Institute." *J. R. Statist. Soc.*, **109**, 325-378.
- MARCH, L. (1926). "Observations sur la méthode représentative et sur le projet de rapport relatif à cette méthode." *Bull. Inst. Int. Statist.*, **22**, 444-451.
- MAYET, P. (1918). "Stichprobenerhebung in der Zwischenzeit zwischen grossen Vollerhebungen längerer Periodizität." *Bull. Inst. Int. Statist.*, **14**, 258-276.
- NEYMAN, J. (1934). "On the two different aspects of the representative method: the method of stratified sampling and the method of purposive selection." *J. R. Statist. Soc.*, **97**, 558-606.
- (1938). "Contributions to the theory of sampling human populations." *J. Amer. Statist. Ass.*, **33**, 101-116.
- NYBOLLE, H. C. (1923). "On Middelfejlen ved partielle Undersogelser." *Nordisk Statistisk Tidsskrift*, **2**, 435-437.
- PIETRA, G. (1944). "La statistique méthodologique italienne de 1939 à 1942." *Rev. Inst. Int. Statist.*, 36-45.
- SETTY, S. P. W. (1941). "An experimental study of some of the methods of representative sampling" (Abstract). *Sankhya*, **5**, 237.

- SMITH, H. F. (1947). "Standard errors of means in sampling surveys with two-stage sampling." *J. R. Statist. Soc.*, **110**, 257-259.
- SNEDECOR, G. W. (1934). "The method of expected numbers for tables of multiple classification with disproportionate sub-class numbers." *J. Amer. Statist. Ass.*, **29**, 389-393.
- and COX, G. M. (1935). "Disproportionate sub-class numbers in tables of multiple classification." *Iowa Agric. Exp. Sta. Res. Bull.* 180.
- STEPHAN, F. F. (1940). See DEMING, W. E.
- (1948). "History of the uses of modern sampling procedures." *J. Amer. Statist. Ass.*, **43**, 12-39.
- STEVENS, W. L. (1948). "Statistical analysis of a non-orthogonal tri-factorial experiment." *Biometrika*, **35**, 346-367.
- STOCK, J. S. and FRANKEL, L. R. (1939). "The allocation of sampling among several strata." *Ann. Math. Statist.*, **10**, 288-293.
- STUART, C. A. V. (1926). "Note sur l'application de la méthode représentative." *Bull. Inst. Int. Statist.*, **22**, 440-443.
- SUKHATME, P. V. (1935). "Contribution to the theory of the representative method." *J. R. Statist. Soc., Suppl.*, **2**, 253-268.
- TORNGVIST, L. (1948). "An attempt to analyse the problem of an economical production of statistical data." *Nordisk Tidsskrift for Teknisk Ökonomi*, **37**, 265-274.
- YATES, F. (1934). "The analysis of multiple classifications with unequal numbers in the different classes." *J. Amer. Statist. Ass.*, **29**, 51-66.
- (1935). "Some examples of biased sampling." *Ann. Eugen.*, **6**, 202-213.
- (1946). "A review of recent statistical developments in sampling and sampling surveys." *J. R. Statist. Soc.*, **109**, 12-43.
- (1948). "The influence of agricultural research statistics on the development of sampling theory" (Abstract). *Econometrica*, **16**, 69-71.
- (1948). "Systematic sampling." *Roy. Soc. Phil. Trans., A*, **241**, 345-377.

B. MACHINE METHODS

- BENJAMIN, K. (1945). "An I.B.M. technique for the computation of ΣX^2 and ΣXY ." *Psychometrika*, **10**, 61-67.
- (1947). "Problems of multiple-punching with Hollerith machines." *J. Amer. Statist. Ass.*, **42**, 46-71.
- BERKSON, J. (1941). "A punch card designed to contain written data and coding." *J. Amer. Statist. Ass.*, **36**, 535-538.
- BLACK, B. J. and OLDS, E. B. (1946). "A punched card method for presenting, analysing and comparing many series of statistics for areas." *J. Amer. Statist. Ass.*, **41**, 347-355.
- CARVER, H. C. (1934). "Punched card systems and statistics." *Ann. Math. Statist.*, **5**, 153-160.
- COMRIE, L. J. (1933). "The Hollerith and Powers tabulating machines." (Printed for private circulation).
- , HEY, G. B. and HUDSON, H. G. (1937). "The application of Hollerith equipment to an agricultural investigation." *J. R. Statist. Soc., Suppl.*, **4**, 210-224.
- (1946). "Application of commercial calculating machines to scientific computing." *Math. Tables and other Aids to Computation*, **2**, 149-159.
- DEMING, W. E., TEPPING, B. J. and GEOFFREY, L. (1942). "Errors in card punching." *J. Amer. Statist. Ass.*, **37**, 525-536.
- DUNN, H. L. and TOWNSEND, L. (1935). "Application of punched card methods to hospital statistics." *J. Amer. Statist. Ass., Suppl.*, **30**, 244-248.
- GEOFFREY, L. (1942). See DEMING, W. E.
- HARTLEY, H. O. (1946). "The application of some commercial calculating machines to certain statistical calculations." *J. R. Statist. Soc., Suppl.*, **8**, 154-183.
- HEY, G. B. (1937). See COMRIE, L. J.
- HUDSON, H. G. (1937). See COMRIE, L. J.
- JOLLIFFE, E. T. (1941). "Fundamental principles in tabulating machine methods of statistical analysis." *J. Exp. Educ.*, **9**, 254-274.

- KEMPTHORNE, O. (1946). "The use of punched card systems for the analysis of survey data, with special reference to the analysis of the National Farm Survey." *J. R. Statist. Soc.*, **109**, 284-295.
- KIMBALL, E. (?). "A fundamental punched card method for technical computations." Bureau of the Census, Washington, D.C. (Undated).
- (?). "A method of technical computations by punched card equipment." Bureau of the Census, Washington, D.C. (Undated).
- MANDEVILLE, J. P. (1946). "Improvements in methods of census and survey analysis." *J. R. Statist. Soc.*, **109**, 111-129.
- McKAY, A. T. (1934). "A new method of handling statistical data." *J. R. Statist. Soc., Suppl.*, **1**, 62-75.
- OLDS, E. B. (1946). See BLACK, B. J.
- TEPPING, B. J. (1942). See DEMING, W. E.
- TODD, H. A. C. (1941). "A note on systematic coding for card sorting systems." *J. R. Statist. Soc., Suppl.*, **7**, 151-154.
- TOOPS, H. A. (1938). "Hollerith Coding." Ohio College Ass. Bull. No. 110, 2269-2274.
- TOWNSEND, L. (1935). See DUNN, H. L.
- VICKERY, C. W. (1938). "Punched card technique for the correction of bias in sampling." *J. Amer. Statist. Ass.*, **33**, 552-556.
- (1939). "On drawing a random sample from a set of punched cards." *J. R. Statist. Soc., Suppl.*, **6**, 62-66.
- WISHART, J. (1947). "Statistical aspects of demobilisation in the Royal Navy." *J. R. Statist. Soc.*, **110**, 27-50.

C. POPULATION CENSUSES

- BLYTHE, R. H. (1947). See JESSEN, R. J.
- COATS, R. H. (1931). "Enumeration and sampling in the field of the census." *J. Amer. Statist. Ass.*, **26**, 270-284.
- DEMING, W. E. (1940). "Sampling problems of the 1940 Census." Cowles Commission for research: Economic Report of 6th Ann. Res. Conf. on Economics and Statistics, Colorado Springs.
- (1940). See STEPHAN, F. F.
- and GEOFFREY, L. (1941). "On sample inspection in the processing of census returns." *J. Amer. Statist. Ass.*, **36**, 351-360.
- and STEPHAN, F. F. (1941). "On the interpretation of censuses as samples." *J. Amer. Statist. Ass.*, **36**, 45-49.
- (1941). See STEPHAN, F. F.
- (1947). See JESSEN, R. J.
- GEOFFREY, L. (1941). See DEMING, W. E.
- GEORGE, R. F. (1936). "A sample investigation of the 1931 Population Census with reference to earners and non-earners." *J. R. Statist. Soc.*, **99**, 147-161.
- GINI, C. (1928). "Une application de la méthode représentative aux matériaux du dernier recensement de la population italienne." *Bull. Inst. Int. Statist.*, **23**, (2), 198-215.
- HANSEN, M. H. (1940). See STEPHAN, F. F.
- (1941). See STEPHAN, F. F.
- (1944). "Census to sample population growth." *Domestic Commerce*, **32**, 6.
- (1944). See HAUSER, P. M.
- and HURWITZ, W. N. (1946). "Sampling methods applied to census work." U.S. Bureau of Census.
- HAUSER, P. M. (1941). "The use of sampling in the census." *J. Amer. Statist. Ass.*, **36**, 369-375.
- (1942). "Proposed annual sample census of population." *J. Amer. Statist. Ass.*, **37**, 81-88.
- and HANSEN, M. H. (1944). "Sample surveys in census work." U.S. Bureau of Census, Dept. of Commerce.
- HURWITZ, W. N. (1946). See HANSEN, M. H.
- JESSEN, R. J., BLYTHE, R. H., KEMPTHORNE, O. and DEMING, W. E. (1947). "On a population sample for Greece." *J. Amer. Statist. Ass.*, **42**, 357-384.

- KAMEDA, T. (1930). "Application of the method of sampling to the first Japanese population census." *Bull. Inst. Int. Statist.*, **25**, (2), 121-132.
- KEMPTHORNE, O. (1947). See JESSEN, R. J.
- STEPHAN, F. F., DEMING, W. E. and HANSEN, M. H. (1940). "The sampling procedure of the 1940 population census." *J. Amer. Statist. Ass.*, **35**, 615-630.
- , ———, ———, ———, (1941). "On the sampling methods in the 1940 population Census." U.S. Dept. of Commerce Census.
- (1941). See DEMING, W. E.
- U.S. BUREAU OF CENSUS (1943). "A revised sample for current surveys." Bureau of Census.

D. SOCIOLOGY, NUTRITION, HEALTH, ETC.

- ANONYMOUS (1941). "Cost of living of the working classes." *J. R. Statist. Soc.*, **104**, 53-58.
- BOSE, A. N. (1941). "Some problems of field operations in labour enquiries." *Sankhya*, **5**, 229-230.
- BOWLEY, A. L. (1913). "Working class households in Reading." *J. R. Statist. Soc.*, **76**, 672-701.
- (1936). "The application of sampling to economic and sociological problems." *J. Amer. Statist. Ass.*, **31**, 474-480.
- BOX, K. and THOMAS, G. (1944). "The Wartime Social Survey." *J. R. Statist. Soc.*, **107**, 151-189.
- CORMICK, T. C. (1937). "Sampling theory in sociological research." *Social Forces*, **16**, 67-74.
- CRUICKSHANK, E. W. H. (1947). "A survey of diets in the maternity wards of Scottish hospitals." *Proc. Nutrition Soc.*, **5**, 149-159.
- DEMING, W. E. (1943). See HANSEN, M. H.
- (1943). See TEPPING, B. J.
- (1944). "On errors in surveys." *Amer. Social Rev.*, **9**, 359-369.
- GHOSH, A. (1946). See MAHALANOBIS, P. C.
- GOODMAN, R. (1947). "Sampling for the 1947 survey of consumer finances." *J. Amer. Statist. Ass.*, **42**, 439-448.
- HANSEN, M. H. and HURWITZ, W. N. (1942). "Relative efficiencies of various sampling units in population inquiries." *J. Amer. Statist. Ass.*, **37**, 89-94.
- and DEMING, W. E. (1943). "On some census aids to sampling." *J. Amer. Statist. Ass.*, **38**, 353-357.
- and HURWITZ, W. N. (1946). "The problems of non-response in sample surveys." *J. Amer. Statist. Ass.*, **41**, 517-529.
- HILTON, J. (1928). "Some further enquiries by sample." *J. R. Statist. Soc.*, **91**, 519-540.
- HURWITZ, W. N. (1942). See HANSEN, M. H.
- (1943). See TEPPING, B. J.
- (1946). See HANSEN, M. H.
- KISER, C. V. (1934). "Pitfalls in sampling for population study." *J. Amer. Statist. Ass.*, **29**, 250-256.
- MAHALANOBIS, P. C., MUKHERJEA, R. and GHOSH, A. (1946). "A sample survey of after-effects of the Bengal famine of 1943." *Sankhya*, **7**, 337-400.
- MASSEY, P. (1942). "The expenditure of 1,360 British middle-class households in 1938-39." *J. R. Statist. Soc.*, **105**, 159-196.
- MATHEN, K. K. (1948). "Studies on the sampling procedure for a general health survey." *Calcutta Statist. Ass. Bull.* No. 3, 106-113.
- McNEMAR, Q. (1940). "Sampling in psychological research." *Psychol. Bull.*, **37**, 331-365.
- MUKHERJEA, R. (1946). See MAHALANOBIS, P. C.
- NEYMAN, J. (1933). "An outline of the theory and practice of representative method applied in social research." Inst. Social Problems, Warsaw. (Polish with an English summary).
- NUTRITION SOCIETY (1945). "Budgeting and dietary surveys of families and individuals." (Abstracts of series of 13 papers read at the 18th and 21st meetings of the Society, 15 Feb. 1944 and 20 May 1944.) *Proc. Nutrition Soc.*, **3**, 1-52 and 110-154.

- PALMER, G. L. (1943). "Factors in the variability of response in enumerative surveys." *J. Amer. Statist. Ass.*, **38**, 143-152.
- PARTEW, M. (1937). See SCHOENBERG, E.
- PIEKALKIEWICZ, J. (1934). "Rapport sur les recherches concernant la structure de la population ouvrière en Pologne selon la méthode représentative." Inst. Social Problems, Warsaw.
- PIETRA, G. (1926). "A particular case of non-representative sampling." *J. Amer. Statist. Ass.*, **21**, 330-332.
- SCHOENBERG, E. and PARTEW, M. (1937). "Methods and problems of sampling presented by the urban study of consumer purchases." *J. Amer. Statist. Ass.*, **32**, 311-322.
- SNEDECOR, G. W. (1939). "Design of sampling experiments in the social sciences." *J. Farm. Econ.*, **21**, 846-855.
- STEPHAN, F. F. (1936). "Practical problems of sampling procedure." *Amer. Social Rev.*, **1**, 569-580.
- (1939). "Representative sampling in large scale surveys." *J. Amer. Statist. Ass.*, **34**, 343-352.
- TEPPING, B. J., HURWITZ, W. H. and DEMING, W. E. (1943). "On the efficiency of deep stratification in block sampling." *J. Amer. Statist. Ass.*, **38**, 93-100.
- THOMAS, G. (1944). See BOX, K.
- U.S. PUBLIC HEALTH SERVICE (1938). "The National Health Survey, preliminary reports. Significance, scope and method." U.S. Public Health Service, Washington, D.C.
- WOODBURY, R. M. (1940). "Methods of family living studies." Int. Labour Office, Series N (Statistics), No. 23.

E. OPINION SURVEYS AND MARKET RESEARCH

- BEVIS, J. C. (1945). "Management of field staffs in the opinion research field." *J. Amer. Statist. Ass.*, **40**, 245-246.
- BLANKENSHIP, A. (1942). "Psychological difficulties in measuring consumer preference." *J. Marketing*, **6**, 66-75.
- CROSSLEY, A. M. (1941). "Theory and application of representative sampling as applied to marketing." *J. Marketing*, **5**, 456-461.
- ECKLER, A. R. and STAUDT, E. P. (1943). "Marketing and sampling uses of population and housing data." *J. Amer. Statist. Ass.*, **38**, 87-92.
- GALLUP, G. (1938). "Government and the sampling referendum." *J. Amer. Statist. Ass.*, **33**, 131-142.
- HANSEN, M. H. (1944). See HAUSER, P. M.
- and HAUSER, P. M. (1945). "Area sampling—some principles of sampling design." *Pub. Opin. Quart.*, **9**, 183-193.
- HAUSER, P. M. and HANSEN, M. H. (1944). "On sampling in market surveys." *J. Marketing*, **9**, 26-31.
- (1945). See HANSEN, M. H.
- HILGARD, E. R. and PAYNE, S. L. (1944). "Those not at home: riddle for pollsters." *Pub. Opin. Quart.*, **8**, 254-261.
- MCPEAK, W. (1945). "Problems of field management in army opinion research." *J. Amer. Statist. Ass.*, **40**, 247-248.
- NORDIN, J. A. (1944). "Determining sample size." *J. Amer. Statist. Ass.*, **39**, 497-506.
- PAYNE, S. L. (1944). See HILGARD, E. R.
- ROPER, E. (1940). "Sampling public opinion." *J. Amer. Statist. Ass.*, **35**, 325-334.
- SILVEY, R. J. E. (1944). "Methods of listener research employed by the British Broadcasting Corporation." *J. R. Statist. Soc.*, **107**, 190-230.
- STAUDT, E. P. (1943). See ECKLER, A. R.
- STEPHAN, F. F. (1941). "Stratification in representative sampling." *J. Marketing*, **6**, 38-46.
- WILKS, S. S. (1940). "Representative sampling and poll reliability." *Pub. Opin. Quart.*, **4**, 261-269.

F. ECONOMICS: SURVEYS OF INDUSTRY, CENSUSES OF PRODUCTION, LABOUR FORCE, ETC.

- BECKNELL, H. E. (1944). "Organising statistical work on a functional basis." *J. Amer. Statist. Ass.*, **39**, 297-302.
- DEDRICK, C. A. and HANSEN, M. H. (1938). "Final report on total and partial unemployment." Vol. IV. The enumerative check census. U.S. Govt. Printing Office.
- ECKLER, A. R. (1945). "The revised census series of current employment estimates." *J. Amer. Statist. Ass.*, **40**, 187-196.
- (1945). "Management of field work and collection of statistics of the labor force." *J. Amer. Statist. Ass.*, **40**, 249-250.
- FRANKEL, L. R. and STOCK, J. S. (1942). "On the sample survey of unemployment." *J. Amer. Statist. Ass.*, **37**, 77-80.
- FRIEDMAN, M. (1936). See KNEELAND, H.
- GEARY, R. C. (1925). "Methods of sampling applied to Irish statistics." *J. Statist. Social Inquiry Soc. Ireland*, **15**, 61-80.
- GURNEY, M. (1946). See HANSEN, M. H.
- HANSEN, M. H. (1938). See DEDRICK, C. A.
- and HURWITZ, W. N. (1944). "A new sample of the population: sampling principles introduced into the Bureau's monthly reports on the labor force." U.S. Bureau of Census.
- and GURNEY, M. (1946). "Problems and methods of the sample survey of business." *J. Amer. Statist. Ass.*, **41**, 173-189.
- HURWITZ, W. N. (1944). See HANSEN, M. H.
- (1946). See HANSEN, M. H.
- KEYFITZ, N. (1945). "The sampling approach to economic data." *Canadian J. Econ. Polit. Sci.*, **11**, 467-477.
- KNEELAND, H., SCHOENBERG, E. H., and FRIEDMAN, M. (1936). "Plans for a study of the consumption of goods and services by American families." *J. Amer. Statist. Ass.*, **31**, 135-141.
- SCHOENBERG, E. H. (1936). See KNEELAND, H.
- STOCK, J. S. (1942). See FRANKEL, L. R.

G. AGRICULTURAL ECONOMICS AND FARM PRACTICE

- ANDERSON, O. (1934). "Description of the Bulgarian agricultural sample census." Bulletin de Statistique, No. 8, Dir. Gen. de Statistique de Bulgarie. (Summary in French).
- BECKER, J. A. and HARLAN, C. L. (1939). "Developments in crop and livestock reporting since 1920." *J. Farm. Econ.*, **21**, 799-827.
- BENEDICT, M. R. (1939). "Development of agricultural statistics in the Bureau of the Census." *J. Farm. Econ.*, **21**, 735-760.
- BEYLEVELD, A. J. (1929). "Determination of a precise indication of change in crop acreage." *J. Amer. Statist. Ass.*, **24**, 405-411.
- BOYD, D. A. (1944). See YATES, F.
- DIMITROFF, St. (1931). "Essai d'élaboration des matériaux du recensement agricole bulgare du 31 Décembre 1926 par l'application de la méthode représentative." *Bull. Inst. Int. Statist.*, **26** (2), 799-817.
- FINKER, A. L., MORGAN, J. J. and MONROE, R. J. (1943). "Methods of estimating farm employment from sample data in North Carolina." N. Carolina Agric. Exp. Sta. Tech. Bull. 75.
- GOODSELL, W. D., JESSEN, R. J., and WILCOX, W. W. (1940). "Procedures which increase the usefulness of farm management research." *J. Farm. Econ.*, **22**, 753-761.
- HARLAN, C. L. (1939). See BECKER, J. A.
- HENDRICKS, W. A. (1944). "The relative efficiencies of groups of farms as sampling units." *J. Amer. Statist. Ass.*, **39**, 367-376.
- HOLMES, I. (1939). "Results of four methods of sampling individual farms." *J. Farm. Econ.*, **21**, 365-374.

- HOLMES, I. (1939). "Research in sample farm census methodology. Part I. Comparative statistical efficiency of sampling units smaller than the minor Civil Division for estimating year to year change." U.S. Dept. Agric., Agric. Marketing Service.
- (1943). "Some sampling uses of data from the census of agriculture." *J. Amer. Statist. Ass.*, **38**, 78-86.
- (1944). "Value of farm products by colour and tenure of farm operator." (Includes description of sample and notes on precision of sample estimates). U.S. Bureau of Census.
- HOPKINS, J. A. (1942). "Statistical comparisons of record-keeping farms and a random sample of Iowa farms for 1939." Iowa Agric. Exp. Sta. Res. Bull. 308.
- HOUSEMAN, E. E. (1944). See JESSEN, R. J.
- (?). "Some developments in sampling and the design of a general purpose sample for enumerative surveys in B.A.E." U.S. Dept. Agric., Bur. Agric. Econ.
- JESSEN, R. J. (1939). "An experiment in the design of agricultural surveys." *J. Farm. Econ.*, **21**, 856-863.
- (1940). See GOODSELL, W. D.
- (1942). "Statistical investigation of a sample survey for obtaining farm facts." Iowa Agric. Exp. Sta. Res. Bull. 304.
- (1943). See STRAND, N. V.
- and HOUSEMAN, E. E. (1944). "Statistical investigations of farm sample surveys taken in Iowa, Florida, and California." Iowa Agric. Exp. Sta. Res. Bull. 329.
- (1945). See KING, A. J.
- (1947). "The master sample project and its use in agricultural economics." *J. Farm. Econ.*, **29**, 531-540.
- KING, A. J. and SIMPSON, G. D. (1940). "New developments in agricultural sampling." *J. Farm. Econ.*, **22**, 341-349.
- and McCARTY, D. E. (1941). "Application of sampling to agricultural statistics with emphasis on stratified samples." *J. Marketing*, **5**, 462-474.
- (1942). See SNEDECOR, G. W.
- and JESSEN, R. J. (1945). "The master sample of agriculture." *J. Amer. Statist. Ass.*, **40**, 38-56.
- MATHISON, I. (1944). See YATES, F.
- McCARTY, D. E. (1941). See KING, A. J.
- MINISTRY OF AGRICULTURE (1946). "National Farm Survey of England and Wales." H.M.S.O.
- MONROE, R. J. (1943). See FINKNER, A. L.
- MORGAN, J. J. (1943). See FINKNER, A. L.
- NAGASAWA, R. (1930). "The method of statistical investigation concerning agricultural production in Japan." *Bull. Inst. Int. Statist.*, **25** (2), 149-178.
- SABIN, A. R. (1940). "A new technique for the estimation of changes in farm employment." (No. 1 of a series of analyses of sample farm data). U.S. Dept. Agric., Agric. Marketing Service.
- SARLE, C. F. (1938). "Methods in sample census research." *J. Farm. Econ.*, **20**, 669-672.
- (1939). "Development of partial and sample census methods." *J. Farm. Econ.*, **21**, 356-364.
- (1939). "Future improvements in agricultural statistics." *J. Farm. Econ.*, **21**, 838-845.
- (1940). "The possibilities and limitations of objective sampling in strengthening agricultural statistics." *Econometrica*, **8**, 45-61.
- SCHUTZ, H. H. (1937). See SHEPARD, J. B.
- SHEPARD, J. B., and SCHUTZ, H. H. (1937). "Selection of areas for sample agricultural enumerations." *J. Farm. Econ.*, **19**, 454-469.
- SIMPSON, G. D. (1940). See KING, A. J.
- SNEDECOR, G. W. and KING, A. J. (1942). "Recent developments in sampling for agricultural statistics." *J. Amer. Statist. Ass.*, **37**, 95-102.
- (1947). "An experiment in the collection of morbidity and mortality data on farm animals." *Proc. U.S. Livestock Sanitary Ass.* (51st meeting), 218-225.
- SOCIAL SCIENCE RESEARCH COUNCIL (1937). "The census of agriculture." Bull. 40. 230, Park Ave., N.Y.

- STRAND, N. V. and JESSEN, R. J. (1943). "Some investigations on the suitability of the township as a unit for sampling Iowa agriculture." Iowa Agric. Exp. Sta. Res. Bull. 315.
- TAEUBER, C. (1944). See TOLLEY, H. R.
- TOLLEY, H. R. and TAEUBER, C. (1944). "Wartime developments in agricultural statistics." *J. Amer. Statist. Ass.*, **39**, 411-427.
- U.S. BUREAU OF AGRICULTURAL ECONOMICS (1936). "Proceedings of conferences on statistical methods of sampling agricultural data." Bur. Agric. Econ., Washington, D.C.
- WILCOX, W. W. (1940). See GOODSELL, W. D.
- YATES, F. (1943). "Methods and purposes of agricultural surveys." *J. R. Soc. Arts*, **91**, 367-379.
- , BOYD, D. A. and MATHISON, I. (1944). "The manuring of farm crops; some results of a survey of fertilizer practice in England." *Emp. J. Exp. Agric.*, **12**, 163-176.

H. CROP ESTIMATION, FORECASTING, ETC.

- BARNARD, M. M. (1936). "An examination of the sampling observations on wheat of the crop-weather scheme." *J. Agric. Sci.*, **26**, 456-487.
- BREWBAKER, H. E. and BUSH, H. L. (1942). "Pre-harvest estimate of yield and sugar percentage based on random sampling technique." Ann. Amer. Soc. Sugar Beet Tech.
- BUCKER, M. I. (?). See PETERS, J. H.
- BUSH, H. L. (1942). See BREWBAKER, H. E.
- CLAPHAM, A. R. (1929). "The estimation of yield in cereal crops by sampling methods." *J. Agric. Sci.*, **19**, 214-235.
- (1929). See WISHART, J.
- (1931). "Studies in sampling technique: cereal experiments. I. Field technique." *J. Agric. Sci.*, **21**, 366-371.
- (1931). "Studies in sampling technique. III. Results and discussion." *J. Agric. Sci.*, **21**, 376-390.
- COCHRAN, W. G. and WATSON, D. J. (1936). "An experiment on observer's bias in the selection of shoot heights." *Emp. J. Exp. Agric.*, **4**, 69-76.
- (1938). See IRWIN, J. O.
- (1940). "The estimation of the yields of cereal experiments by sampling for the ratio of grain to total produce." *J. Agric. Sci.*, **30**, 262-275.
- DHANNALAL (1940). See KALAMKAR, R. J.
- EDGAR, J. L. and DE WET, A. F. (1935). "An experiment in sampling technique for size and colour of apples: Bramley's Seedling on Rootstock No. 11, 1934 crop." Rep. E. Malling Res. Sta., 130-4.
- (1938). See HOBLYN, T. N.
- FEDERER, W. T. (1946). See HOUSEMAN, E. E.
- GHOSH, B. (1947). "Crop estimation in India: a brief review of the sampling methods." Calcutta Statist. Ass. Bull., No. 1, 5-12.
- HEATH, O. V. S. (1934). "Sampling and growth observations in plant development studies in cotton." Report of Proceedings of Empire Cotton Growing Corporation: 2nd conference on cotton growing problems; 96-110.
- HENDRICKS, W. A. (1942). "Theoretical aspects of the use of the crop meter." No. 2 of a series of analyses of sample farm data. U.S. Dept. Agric., Agric. Marketing Service.
- HOBLYN, T. N. and EDGAR, J. L. (1938). "Experiments in sampling techniques. II. Size and colour of Allington Pippin, 1936 crop." Rep. E. Malling Res. Sta. for 1937, 168-172.
- HOUSEMAN, E. E., WEBER, C. R., and FEDERER, W. T. (1946). "Pre-harvest sampling of soybeans for yield and quality." Iowa Agric. Exp. Sta. Res. Bull. 341.
- HUBBACK, J. A. (1927). "Sampling for rice yield in Bihar and Orissa." Imp. Agric. Res. Inst., Pusa, Bull. 166. (Reprinted 1946: *Sankhya*, **7**, 281-294).
- HUDSON, H. G. (1939). "Population studies with wheat. I. Sampling." *J. Agric. Sci.*, **29**, 76-110.
- IMMER, F. R. (1932). "A study in sampling technique with sugar beets." *J. Agric. Res.*, **44**, 633-647.

- IRWIN, J. O., COCHRAN, W. G., and WISHART, J. (1938). "Crop estimation and its relation to agricultural meteorology." *J. R. Statist. Soc., Suppl.*, **5**, 1-45.
- JEBE, E. H. (1940). See KING, A. J.
- KALAMKAR, R. J. (1932). "A study in sampling techniques with wheat." *J. Agric. Sci.*, **22**, 783-796.
- and DHANNALAL (1940). "Sampling studies in cotton varietal trial." *Sankhya*, **4**, 567-576.
- (1941). "A note on crop-cutting experiments on cotton." *Sankhya*, **5**, 345-348.
- (1944). See PANSE, V. G.
- (1944). See PANSE, V. G.
- (1945). See PANSE, V. G.
- KING, A. J. and JEBE, E. H. (1940). "An experiment in pre-harvest sampling of wheat fields." *Iowa Agric. Exp. Sta. Res. Bull.* 273.
- , McCARTY, D. E. and McPEAK, M. (1942). "An objective method of sampling wheat fields to estimate production and quality of wheat." U.S. Dept. Agric. Tech. Bull. 814.
- LEGATT, C. W. (1941). "A study of the relative efficiency of seed sampling methods." *Can. J. Res.*, **19**, 156-162.
- MAHALANOBIS, P. C. (1940). "A sample survey of the acreage under jute in Bengal." *Sankhya*, **4**, 511-530.
- (1945). "Report on the Bihar crop survey: Rabi season, 1943-4." *Sankhya*, **7**, 29-106.
- (1946). "Sample surveys of crop yields in India." *Sankhya*, **7**, 269-280.
- MALLIK, A. K., SATAKOPAN, V. and RAO, S. G. (1945). "A study of the estimation of the yield of wheat by sampling." *Indian J. Agric. Sci.*, **15**, 219-225.
- MCCANDLISS, D. A. (1941). "Objective sampling in estimating southern crops." *J. Farm. Econ.*, **23**, 246-255.
- MCCARTY, D. E. (1942). See KING, A. J.
- McPEAK, M. (1942). See KING, A. J.
- PANSE, V. G. (1938). "Preliminary studies on sampling in field experiments." *Sankhya*, **4**, 139-148.
- and KALAMKAR, R. J. (1944). "Forecasting and estimation of crop yields." *Current Science*, **13**, 120-124.
- and KALAMKAR, R. J. (1944). "A further note on the estimation of crop yields." *Current Science*, **13**, 223-225.
- and KALAMKAR, R. J. (1945). "A large-scale yield survey on cotton." *Current Science*, **14**, 287-291.
- (1946). "Plot size in yield surveys on cotton." *Current Science*, **15**, 218-219.
- PEARCE, S. C. (1943). "An investigation into means of reducing the labour needed for recording crops." Rep. E. Malling Res. Sta. for 1942, 36-40.
- (1944). "Sampling methods for the measurement of fruit crops." *J. R. Statist. Soc.*, **107**, 117-126.
- PETERS, J. H. and BUCKER, M. I. (?) "The 1940 section survey of crop acreages in Indiana and Iowa." U.S. Dept. Agric., Bur. Agric. Econ.
- RAO, S. G. (1945). See MALLIK, A. K.
- SARLE, C. F. (1932). "Adequacy and reliability of crop yields estimates." U.S. Dept. Agric. Tech. Bull. 311.
- SATAKOPAN, V. (1945). See MALLIK, A. K.
- SREENIVASAN, P. S. (1943). "Studies on the estimation of growth and yield of jowar by sampling." *Indian J. Agric. Sci.*, **13**, 399-412.
- SUKHATME, P. V. (1945). "Random sampling for estimating rice yields in Madras Province." *Indian J. Agric. Sci.*, **15**, 308-318.
- (1946). "Bias in the use of small-size plots in sample surveys for yield." *Nature*, **157**, 630.
- (1946). "Size of sampling unit in yield surveys." *Nature*, **158**, 345.
- (1947). "The problem of plot size in the large scale yield surveys." *J. Amer. Statist. Ass.*, **42**, 297-310.
- (1947). "Use of small size plots in yield surveys." *Nature*, **160**, 542.
- WATSON, D. J. (1934). See YATES, F.
- (1936). See COCHRAN, W. G.
- (1937). "The estimation of leaf areas." *J. Agric. Sci.*, **27**, 474-483.

- WEBER, C. R. (1946). See HOUSEMAN, E. E.
 WET, A. F. DE (1935). See EDGAR, J. L.
 WISHART, J. and CLAPHAM, A. R. (1929). "A study in sampling technique. The effect of artificial fertilizers on the yield of potatoes." *J. Agric. Sci.*, **19**, 600-618.
 ——— (1938). See IRWIN, J. O.
 YATES, F. and WATSON, D. J. (1934). "Observer's bias in sampling observations on wheat." *Emp. J. Exp. Agric.*, **2**, 174-177.
 ——— and ZACOPANAY, I. (1935). "The estimation of the efficiency of sampling, with special reference to sampling for yield in cereal experiments." *J. Agric. Sci.*, **25**, 545-577.
 ——— (1935). "The place of quantitative measurements on plant growth in agricultural meteorology and crop forecasting." Conference of Empire Meteorologists, London, 169-172.
 ——— (1936). "Crop estimation and forecasting: indications of the sampling observations on wheat." *J. Min. Agric.*, **43**, 156-162.
 ——— (1936). "Applications of the sampling technique to crop estimation and forecasting." Manchester Statist. Soc.
 ZACOPANAY, I. (1935). See YATES, F.

I. FORESTRY AND LAND UTILIZATION SURVEYS

- BICKERSTAFF, A. (1947). "One-fifth acre versus one-tenth acre plots in sampling immature stands." Dominion Forest Service, Canada. Silvicultural Research Note 83.
 ——— (1947). "Sampling efficiency of line plot survey on Riding Mountain research area." Dominion Forest Service, Canada. Silvicultural Research Note 84.
 BLYTHE, R. H. (1945). "The economics of sample size applied to the scaling of saw logs." *Biometrics*, **1**, 67-70.
 FINNEY, D. J. (1949). "Random and systematic sampling in timber surveys." *Forestry*, **22**, 64-99.
 FOGH, I. F. (1943). "Sampling methods in log scaling." *For. Chron.*, **19**, 127-138.
 GEVORKIANTZ, S. R. (1934). See MUDGETT, B. D.
 ——— (1939). See GIRARD, J. W.
 GIRARD, J. W. and GEVORKIANTZ, S. R. (1939). "Timber cruising." U.S. Dept. Agric. Forest Service, Washington.
 HASEL, A. A. (1937). "Arrangement of cruise plots to permit a valid estimate of sampling error." Forest and Range Exp. Sta.
 ——— (1938). "Sampling error in timber surveys." *J. Agric. Res.*, **57**, 713-736.
 ——— (1941). "Estimation of vegetation type areas by linear measurement." *J. For.*, **39**, 34-40.
 ——— (1942). "Estimation of volume in timber stands by strip sampling." *Ann. Math. Statist.*, **13**, 179-206.
 HURWITZ, W. N. (1943). "Working plan for annual census of lumber produced in 1943." U.S. Dept. Agric. Forest Service.
 LEXEN, B. (1941). "The application of sampling to log scaling." *J. For.*, **39**, 624-631.
 LOOMIS, R. D. (1946). "Accuracy in timber estimating." *For. Chron.*, **22**, 201-202.
 MUDGETT, B. D. and GEVORKIANTZ, S. R. (1934). "Reliability of forest surveys." *J. Amer. Statist. Ass.*, **29**, 257-281.
 MULLOY, G. A. (1946). See ROBERTSON, W. M.
 OSBORNE, J. G. (1942). "Sampling errors of systematic and random surveys of cover-type areas." *J. Amer. Statist. Ass.*, **37**, 256-264.
 PECHANEC, J. (1941). "Sampling error in range surveys of sagebrush grass vegetation." *J. For.*, **39**, 52-54.
 PROUDFOOT, M. J. (1942). "Sampling with transverse traverse lines." *J. Amer. Statist. Ass.*, **37**, 265-270.
 ROBERTSON, W. M. and MULLOY, G. A. (1946). "Sample plot methods." Dominion Forest Service, Ottawa.
 THOMPSON, A. P. (1945). "A sampling approach to New Zealand timber cruising problems." *N.Z. J. For.*, **5**, 103-117.

J. ESTIMATION OF WILD POPULATIONS

- BEALL, G. (1939). "Methods of estimating the population of insects in a field." *Biometrika*, **30**, 422-439.
- COCHRAN, W. G. (1938). "The information supplied by the sampling results." *Ann. Appl. Biol.*, **25**, 383-389. (Appendix to LADELL, R. S., "Field experiments on the control of wireworms." *Ann. Appl. Biol.*, **25**, 341-382.)
- DELURY, D. B. (1947). "On the estimation of biological populations." *Biometrics*, **3**, 145-167.
- FINNEY, D. J. (1941). "Wireworm populations and their effect on crops." *Ann. Appl. Biol.*, **28**, 282-295.
- (1942). See YATES, F.
- (1946). "Field sampling for the estimation of wireworm populations." *Biometrics*, **2**, 1-7.
- GREENSLADE, R. M. and PEARCE, S. C. (1940). "Field sampling for the comparison of infestations of strawberry crops by the aphid *Capitophorus fragariae* Theob." *Pomology and Hort. Sci.*, **17**, 308-317.
- HOEL, P. G. (1943). "The accuracy of sampling methods in ecology." *Ann. Math. Statist.*, **14**, 289-300.
- JACKSON, C. H. N. (1939). "The analysis of an animal population." *J. Anim. Ecol.*, **8**, 238-246.
- JONES, E. W. (1937). "Practical field methods of sampling soil for wireworms." *J. Agric. Res.*, **54**, 123-134.
- MEYERS, M. T. and PATCH, L. H. (1937). "A statistical study of sampling in field surveys of the fall population of the European Corn Borer." *J. Agric. Res.*, **55**, 849-872.
- MINISTRY OF AGRICULTURE (1944). "Wireworms and food production." A wireworm survey of England and Wales. 1939-1942. Bull. 128. H.M.S.O.
- PATCH, L. H. (1937). See MEYERS, M. T.
- PEARCE, S. C. (1940). See GREENSLADE, R. M.
- WADLEY, F. M. (1945). "An application of the Poisson series to some problems of enumerations." *J. Amer. Statist. Ass.*, **40**, 85-92.
- YATES, F. and FINNEY, D. J. (1942). "Statistical problems in field sampling for wireworms." *Ann. Appl. Biol.*, **29**, 156-167.

INDEX

When a subject is dealt with in the whole of a section only the first page of the section is entered. Relevant examples which occur at the end of such sections are not separately indexed.

- ABSTRACTION** of data, 109, 123.
 Accuracy, 32, 143, 145, 247.
 Addressograph list, 82.
 Administrative areas, sampling by, 36, 75.
 Administrative organization for surveys, 102.
 Advertising, 79.
 Advisory Economists, 59.
 Aerial photographs, 35, 42, 74, 86.
 Agriculture, frames for, 81-87.
 Agricultural Meteorological Committee, 13.
 Allied Mission for Observing the Greek Election, 71.
 Alternative estimates, 145.
 Analysis of results, 108-141; methods, 109; by Cope-Chat cards, 110; by punched cards, 112-123; sampling for, 128; critical, 109, 131; of two-way tables, 131-141; errors in, 124.
 Analysis of variance, 205; interpretation of, 266; applications of, 218, 250, 254, 269, 280, 281; reporting, 143; examples of, 208, 210, 226, 228, 252, 273.
 Animal populations, 44.
 Area sampling, 68-79, 82-87; *see also* systematic sample from areas.
 Areas, measurement by sampling, *see* point and line sampling.
 Areas, selection with equal probability, 82.
 Attributes, *see* qualitative variates.
- BALANCED** differences, 231.
 Balanced sample, 39, 174, 221.
 Bias, 9-17, 143; permissible, 17; estimation of, 239; in selection, 9-15, 65, 80, 84, 165, 222, 240; in demarcation of units, 15, 164; in eye estimates, 43, 88, 163, 165, 222; in estimation, 16, 73, 77, 145, 162, 174; in estimate of error, 198.
 Bibliography, 299-311.
 Biological sampling, 48, 236.
 Blocks, city, 68, 71.
 Blocks, randomized, 105.
 Blythe, R. H., 71.
- Box, K., 55.
 Boyd, D. A., 81.
 British Tabulating Machine Co., *see* punched cards.
 Bureau of Agricultural Economics, 73.
 Bureau of the Census, 73.
- CALCUTTA** Institute of Statistics, 83.
 Calibration of eye estimates, 43, 88, 165, 222.
 Cards, 104, 109; *see also* Cope-Chat cards and punched cards.
 Causal relationship, 131, 188.
 Cells, 41.
 Census of Woodlands, 16, 44, 46, 83, 99, 163, 220, 232, 238, 242, 257, 288.
 Central Office of Information, 78.
 Change, *see* successive occasions.
 Checks on field work, 106; on computations, 124; by comparison with complete returns, 31, 144.
 Chi-squared test, 22, 28, 200.
 Cluster sampling, 20.
 Cochran, W. G., 270, 273.
 Coding, 109, 110, 111, 118-123, 126.
 Coefficient of variation, 184.
 Collator, 118.
 Collective characteristics, 122.
 Commercial undertakings, 79.
 Comparability, 50.
 Complete census, 3; combination with sample, 47.
 Composite sampling scheme, 46.
 Compulsory returns, 59.
 Computations, preliminary, 108, 117, 123; checks on, 124; *see also* analysis.
 Constants, fitting of, 137.
 Consumer preferences, 79.
 Control, machine, 114.
 Control characters, 40.
 Cope-Chat cards, 104, 109, 110.
 Correlation coefficient, 176, 199.
 Costs, 143; minimization of, 283-296.
 Counter, 113.
 Counts, in analysis, 109, 110, 111, 113, 115.
 Covariance, 198.
 Coverage, 49, 141.
 Cox, G., 138.

Crockery breakages, 55.
 Crop estimation, 42, 87-92; areas, 168, 224, 272, 286, 289, 290; yields per acre, 165, 168, 222, 224, 264, 273, 286, 289, 291; sampling of standing crop, 13, 15, 90, 99.
 Crop forecasting, 92.
 Cross footing multiplying punch, 117.
 Crowds, 43.
 Cruising, 42, 90.

DEDUCTIONS from surveys, qualifications of, 131, 212.
 Defective sample, adjustment for, 129.
 Degrees of freedom, 185, 206.
 Deming, W. E., 65, 71, 127.
 Description of survey, 141.
 Distributors, 114.
 Domains of study, 24, 28; estimates for, 146; errors for, 146, 202, 210.
 Duplicate samples, 242.
 Duplication in frame, 60.
 Dwellings, 67, 74.

ECKLER, A. R., 77.
 Economic institutions, frames for, 79.
 Efficiency, 109, 144, 145, 200, 246-296; definition, 247; *see also under types of sample*.
 Efficient estimate, 247.
 Election forecasts, 80.
 Electoral lists, 66, 71.
 Employment, 76.
 End corrections, 175.
 England and Wales, *see* Census of Woodlands, National Farm Survey, Survey of Fertilizer Practice.
 Equipment, 142.
 Error, limits of, 191, 236.
 Error, sampling; *see* random sampling error *and* bias.
 Error graph, 235.
 Errors, in observation and measurement, 15; in fieldwork, 106; in computations, 124; rounding off, 238; grouping, 238; *see also* investigators, tests of.
 Estimates, alternative, 145.
 Estimation of population values, 145-182; rules for, 147; of sampling errors, 183-245; of size of sample and relative efficiency, 94-99, 246-296; *see also under types of sample*.
 Experiments, 105, 131, 212.
 Explanatory notes, 103.
 Exploratory surveys, 48, 99.
 Eye estimates, calibration of, 43, 88, 165, 222.

FACTORIAL design, 105.
 Factories, 79.
 Factors, 131.
 Families, 121, 217.
 Family Census (U.K.), 53, 64, 130.
 Family income (Norfolk-Portsmouth), 96, 186, 239.
 Farms, 74, 81, 273, 288, *see also* Hertfordshire farms.
 Fertilizer Practice, *see* Survey of.
 Fiducial probability, 191, 236.
 Field, of punched card, 113.
 Field work, organization of, 102-106; control of accuracy, 106.
 Fields, 81, 273, 288.
 Finite sampling, corrections for, 187, 246.
 Finney, D. J., 237.
 Fisher, R. A., 105, 201, 267.
 Fitting constants, 137.
 Fixed sample, 45.
 Flats, 68.
 Follow up, *see* non-response.
 Food offices, 64.
 Forecasts, 80, 92.
 Forestry, frames for, 83-87, *see also* Census of Woodlands.
 Forms, 57, 103-105.
 Frame, 20, 60-87, 144; defects of, 60, 144; human populations, 62-78; economic institutions, 79; agriculture, 81-87; forestry, 83-87; construction of second-stage, 34, 68, 71; from censuses, 65; from lists, 63, 66, 70, 79; from maps, 68-75, 79, 81-86; from aerial photographs, 86.
 Frankel, L. R., 77.
 Functions, errors of, 196.

GALVANI, L., 40.
 Gamma function, 293.
 Gang-punching, 116.
 Geoffrey, L., 127.
 Geographical scope, 49, 141.
 Gini, C., 40.
 Glass, D. V., 130.
 Greece, population census, 71.
 Grouping, 118, 186, 188, 238.

HALF-OPEN interval, 67, 68.
 Hansen, M. H., 36, 65.
 Haphazard selection, 10.
 Hertfordshire farms, samples for wheat acreage, 30, 36; random sample, 97, 152, 159, 161, 162, 189, 205, 214, 216, 220, 252, 258; stratified sample, 150, 203, 249, 251; variable sampling fraction, 154, 207, 252, 255, 256;

INDEX

- samples of parishes, 169, 226, 264, 266;
- two-stage samples, 270, 271, 290.
- Hollerith, *see* punched cards.
- Households, 78, 121, 217.
- Houses, percentages defective, 149, 195.
- Hurwitz, W. N., 36.

- I.B.M., *see* punched cards.
- Inaccuracy, in frame, 60.
- Inadequacy, in frame, 60, 61, 78.
- Incomplete census, 2.
- Incomplete coverage, 10.
- Incomplete results, adjustment for, 129.
- Incompleteness, in frame, 60.
- Independence, 196.
- Independent samples, 45.
- Index numbers, calculation of, 117.
- India, Calcutta Institute of Statistics, 83.
- Industrial undertakings, 79.
- Information, description of, 141; required, 51; methods of collection, 57; practicability, 55.
- Insect populations, 44.
- Instructions, 103, 141.
- Integral values of supplementary variate, 217.
- Interactions, 105, 140, 211.
- Interpenetrating samples, 44, 105, 107, 143, 241, 242; examples, 83.
- Inter-relations between units, 54.
- Intra-class correlation, 267.
- Investigators, 58, 105; tests of, 44, 99, 105, 107, 143, 241; instructions to, 103; conditions of work, 106.
- Iowa State College Statistical Laboratory, 73.
- Italy, population census, 40.

- JESSEN, R. J., 71, 73.

- KEMPTHORNE, O., 71, 117.
- King, A. J., 73.
- Kiser, C. V., 11.
- Kraals, 159, 214.

- LAND utilization surveys, 86.
- Limits of error, 191, 236.
- Line sampling, 42, 85, 86; errors, 229; examples, 232.
- Linear functions, standard errors of, 196.
- Listing, 114.
- Lists, *see* frame and systematic sample.
- Livestock, 81.
- Localized population survey, 75.
- Losses due to errors, 292.

- MAHALANOBIS, P. C., 83, 92.
- Maps, use as frames, 68-75, 79, 81-86; areas from, *see* point and line sampling.
- Marginal categories, 49.
- Mark sensing, 109.
- Market research, 79.
- Master cards, 116.
- Master sample, 65, 73, 75.
- Mathison, I., 81.
- Mean, arithmetic, 145; rule for estimation of, 148; geometric, 145; working, 185; correction for, 185.
- Mean square, 206.
- Mean square deviation, 183.
- Measurement, errors in, 15.
- Median, 145.
- Milk, composition of, 178, 181, 234, 235, 262.
- Ministry of Agriculture, 82, 128, 158.
- Ministry of Home Security, 67.
- Morbidity, 12.
- Moving observer, 43.
- Multi-phase sample, 38; estimates, 157, 159, 162; errors, 213, 219; size and efficiency, 258, 286.
- Multi-stage sample, 18, 34; estimates, 170, 171; errors, 226; size and efficiency, 98, 268, 285; examples, 71, 77, 81, 84.
- Multi-stage sample with uniform overall sampling fraction, 36, 148, 171, 278, 287; examples, 71, 77; with adjustment of proportions of second-stage units, 78.
- Multiple classification, analysis of, 131-141.
- Multiple punching, 119.
- Multiple stratification, 25, 254; examples, 73, 77; without control of sub-strata, 25.
- Multiplying punch, 117.

- NATIONAL Agricultural Advisory Service, 59.
- National Farm Survey, 115, 117, 128, 158.
- National Register, 64.
- Natural units, 20; hierarchy of, 121.
- Non-response, 59, 107, 130; sub-sample for, 108.
- Norfolk-Portsmouth, Virginia, 186.

- Normal distribution, 190 ; sample from, 149, 185, 188, 190, 192, 298.
 Normal law of error, 190.
 Notation, 7, 146.
 Nuffield Trust, 78.
- OBSERVATION**, errors of, 15.
 Observers, *see* investigators.
 Opinion surveys, 79 ; effect of stratification, 248.
 Optimal values, 284.
 Optional allocation, 18, 29.
 Ordnance Survey, 75, 82, 83, 84.
 Orthogonality, 211.
 Out-of-date frame, 60.
 Overall estimates, 175.
- PARTIAL** replacement, *see* successive occasions.
 Patterson, H. D., 179, 180.
 Percentage standard deviation, 96, 184.
 Percentage standard error, 95, 184.
 Percentages, choice of, 108 ; calculation of, 117 ; estimation of, 148 ; standard error of, 94, 193.
 Personnel, 142.
 Phase, *see* multi-phase.
 Pilot surveys, 48, 99, 273.
 Planning of surveys, 48-101, 246, 294.
 Point sampling, 35, 69, 82, 86 ; estimates, 167 ; errors, 224 ; size and efficiency, 262, 286 ; examples, 272.
 Pooled estimate of error, 205, 236.
 Pooling of classes, 137.
 Population, human, 121, 217 ; frames for, 62-78 ; localized surveys of, 75-78 ; special classes, 78.
 Population, statistical, 20 ; finite, 187 ; to be covered, 49, 141 ; values, *see* estimation.
 Population census, Greece, 71 ; Italy, 40 ; Southern Rhodesia, 159, 214 ; U.K., 53 ; U.S.A., 65 ; frame from, 65.
 Postal enquiry, 58, 107, 130.
 Potato survey, 131-141, 199, 209.
 Powers-Samas, *see* punched cards.
 Precision, relative, 246-283 ; definition, 247.
 Precoding, 120.
 Preliminary computations, *see* computations.
 Preliminary count (of dwellings), 68.
 Preliminary estimates, 85, 92.
 Printing, 113.
 Probability of selection proportional to size, sample with, 35, 36 ; errors, 224, 225 ; estimates, 167, 169 ; size and efficiency, 262 ; examples, 36, 71, 77 ; *see also* point sampling.
 Progressive digitizing, 115.
 Progressive totals, 115.
 Proportion, *see* percentages.
 Public opinion polls, 79.
 Punched cards, 109, 112-123, 126.
 Punching, 109, 112, 120, 126.
 Purpose of survey, 49, 51, 141.
 Purposive selection, 40, 80, 142.
- QUALITATIVE** variates, 94, 193, 232, 248 ; rule for, 148.
 Quality control, 48.
 Questionnaires, 52-59, 103-105 ; tests of, 99, 104, 105 ; postal, 58, 107.
 Questions, wording of, 103.
 Quota method, 80, 142.
- RAISING** factor, 147 ; overall, 170.
 Random numbers, 21, 297.
 Random sample, 10, 21 ; estimates, 145, 148, 152, 159, 162 ; errors, 183-196, 212, 217, 218 ; size and efficiency, 94, 248, 249, 256 ; examples, 31, 83.
 Random sampling error, 2, 9, 17 ; estimation of, 183-245 ; by sampling, 238 ; from duplicate samples, 242 ; presentation of, 243 ; *see also* under types of sample.
 Random selection, 21 ; examples, 22.
 Random selection from areas, 22.
 Randomized blocks, 105.
 Rating offices, 66.
 Ratio, rule for estimation of, 148 ; standard error of, 198, 212.
 Ratio method, *see* supplementary information.
 Ration books, 64.
 Ratios, calculation of, 117.
 Read, D. R., 299.
 Regression coefficient, 155, 199 ; standard error of, 219 ; *see also* supplementary information and calibration of eye estimates.
 Rents, 158.
 Repeated surveys, 17, 79 ; *see also* successive occasions.
 Reports, 141.
 Representative sample, 9, 84.
 Reproducing punch, 116.
 Response, failure of, *see* non-response.
 Rolling total tabulator, 114.
 Rounding off, 118, 238.
 Rowntree, B. Seebohm, 4.
 Royal Commission on Population, 64, 130.

INDEX

- SAMPLE**, types of, 20-47; *see also under separate types*.
- Sample census, 2.
- Sample survey, 4.
- Sampling error, *see* random sampling error and bias.
- Sampling fraction, 18, 23, 24, 147, 148.
- Sampling process, 1; in censuses and surveys, 2-6; census, incomplete sample, 2; survey, sample, 4.
- Sampling units, 20; choice of, 19; multi-stage, 34; inter-relations between, 54; variation in size of, 19, 98, 279; rule for estimation of number in population, 147.
- Scoring, 147, 148, 195.
- Selection, methods of, 21, 142.
- Selection with probability proportional to size, *see* probability of selection.
- Shaul, J. R. H., 160.
- Sheppard's correction, 239.
- Shops, 79.
- Sickness, 12.
- Size, probability proportional to, *see* probability of selection.
- Size of sample, determination of, 94-101, 246-296; *see also under types of sample*.
- Size of strata, variation in, 98, 280.
- Size of unit, effect on sampling error, 19, 98; variation in, 279.
- Snedecor, G., 105, 138.
- Social Survey, 78.
- Soil analysis, 82.
- Soil temperatures, 282.
- Solids-not-fat, *see* milk.
- Sorter, 113.
- Sorter-counter, 113.
- Sorting, 110.
- Southern Rhodesia, 159, 214.
- Stage, *see* multi-stage.
- Standard deviation, 96, 183, 190; percentage, 96, 184.
- Standard error, 94, 183; percentage, 95, 184; of qualitative variates, 94, 193; of mean, 96, 184, 187; of total, 96, 184, 187; of ratio, 198, 212; of multiple, 196; of product, 198; of sum, 197; of difference, 196; of linear function, 196; of weighted mean, 197; of standard deviation, 192; effect of lack of independence, 198; *see also under types of sample*.
- Standardization, 157, 158, 159, 162, 213, 219.
- Statistical analysis, *see* analysis.
- Statistician, functions of, 6, 49.
- Stephan, F. F., 65.
- Stevens, W. L., 138.
- Stock, J. S., 77.
- Stones, 12.
- Stratification after selection, 25, 32, 152, 205.
- Stratification, multiple, *see* multiple stratification.
- Stratified sample, variation in size of strata, 98, 280.
- Stratified sample with one unit per stratum, 24, 78, 280.
- Stratified sample with uniform sampling fraction, 17, 23, 146; estimates, 150, 160, 164; errors, 201, 205, 215, 221; size and efficiency, 98, 248, 249, 256; examples, 31, 37.
- Stratified sample with variable sampling fraction, *see* variable sampling fraction.
- Streets, sampling by, 67, 69.
- Sub-Commission on Statistical Sampling, 141.
- Sub-sample for non-response, 60; on successive occasions, 45; *see also* multi-phase sample.
- Sub-totals, 115.
- Substitution, 10, 108.
- Successive occasions, sampling on, 17, 45; estimates, 175, 179; errors, 233; size and efficiency, 260; examples, 77.
- Sukhatme, P. V., 15, 255.
- Sum of squares, 185, 206; calculation of, 185.
- Summary punch, 116.
- Supervision, 19, 105.
- Supplementary information, 18, 32, 38, 98, 145, 146; ratio method, 71, 155-162, 171-174, 198, 212-218, 256; regression method, 155, 162-165, 171, 218-222, 256; effects of errors in, 213.
- Survey, definition, 4.
- Survey of Fertilizer Practice, 57, 81, 111, 123, 171, 227, 240, 257, 264, 291.
- Syracuse, U.S.A., 11.
- Systematic sample from areas, 41; estimates, 174; errors, 229; size and efficiency, 282; examples, 83.
- Systematic sample from lists, 10, 29; estimates, 174; errors, 229; examples, 64, 65, 67, 81.
- DISTRIBUTION**, 192.
- Tabulation, machine, 114.
- Tabulator, 113.
- Telephone enquiries, 80.
- Temperatures, soil, 282.
- Tepping, E. J., 127.
- Terminology, 7.
- Tests of questionnaires, 99, 104, 105; of investigators, 99, 105, 241; of significance, 188, 200.
- Thomas, G., 55.
- Timber, *see* Census of Woodlands.

INDEX

- Totals, computation of, 109, 110, 111, 113; rule for estimation of, 147.
 Tracks, 70.
 Trailers, 116, 120.
 Training, 99, 105.
 Transformations, 236.
 Travelling, 19.
 Two-nine feature, 120.
 Two-way tables, analysis of, 131.
- U.S.A.** employment estimates, 76;
 master sample, 73; population census, 65; presidential election, 1948, 80.
 Undeveloped areas, 34, 42, 296; frames for, 70, 85-87.
 Unemployment, 76.
 Uniform sampling fraction, 23; overall, *see* multi-stage sample with uniform overall sampling fraction.
 Unit, natural, *see* natural units.
 Unit, sampling, *see* sampling units.
 United Kingdom, effects of air raids, 67;
 Family Census, 53, 64, 130; localized surveys, 76; Population Census, 53.
 United Nations, 141, 299.
- VARIABLE** sampling fraction, sample with, 18, 28; estimates, 153, 161, 164;
- errors, 201, 205, 216, 221; size and efficiency, 98, 254, 256, 285; examples, 31, 71, 81, 115, 128, 171.
 Variance, 183; unequal, 201, 207.
 Variate, 146.
 Variation, coefficient of, 184.
 Villages, 70.
- WEIGHTED** mean, 17; of sub-class means, 134; of differences of sub-class means, 136; standard error of, 197.
 Weighting factors, 108, 123.
 Wheat, 13, 15, 166, 223, 273; *see also* Hertfordshire farms.
 Wireworms, 236.
 World statistics, 51.
- YATES**, F., 12, 13, 81, 138, 211, 237, 268, 283.
 Yield per acre, *see* crop estimation *and* crop forecasting; bias in, 16.
- ZACOPANAY**, I., 268.